3D Integration

Bob Patti, CTO

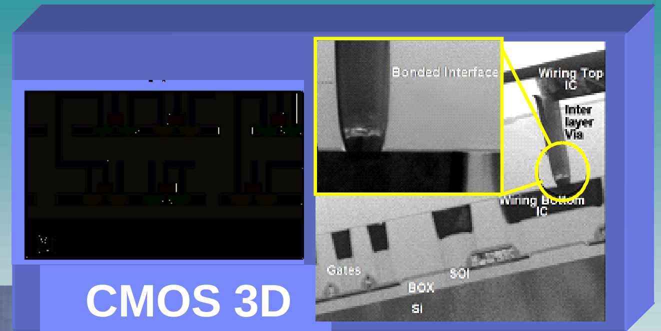
rpatti@tezzaron.con

Evolution of 3D Integration

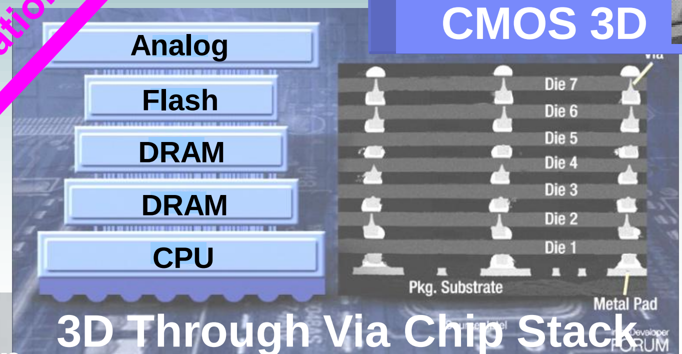
Technology Investment in the Z-Dimension

- **3D technologies continue the sequence of interconnect advances**
- **Return balance to device scaling**
- **Enable new capabilities not available in 2D**

Increasing 3D Integration

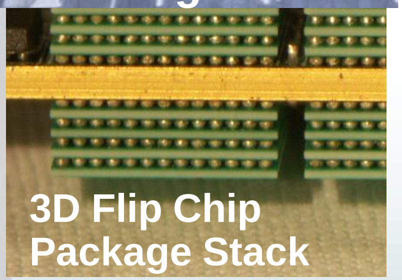


CMOS 3D



3D Through Via Chip Stack

3D Pkg Chip Stack



3D Flip Chip Package Stack

Wire bonded chip stacked 3D

- **3D packaging R&D now pervasive in industry, academia**
- **Through via technology emerging as predominant path**
- **3D has *always* been large volume, but now integrating higher technologies**

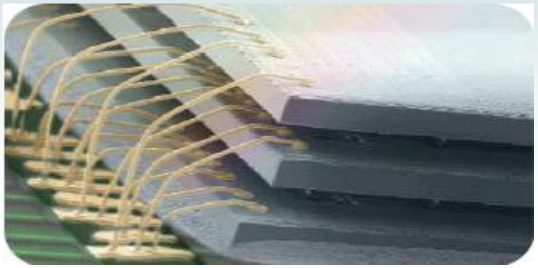
K. Bernstein, “New Dimensions in Performance: Emerging 3D Integration Technologies," VMIC 2006

3D Stacking Approaches

Chip Level

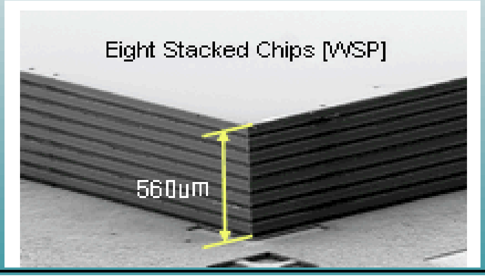
- Ziptronix
- Xan3D
- Vertical Circuits

Amkor : 4S CSP (MCP)



Irvine Sensors : Stacked Flash

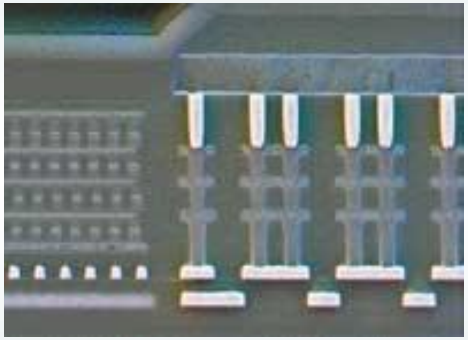
Samsung : Stacked Flash

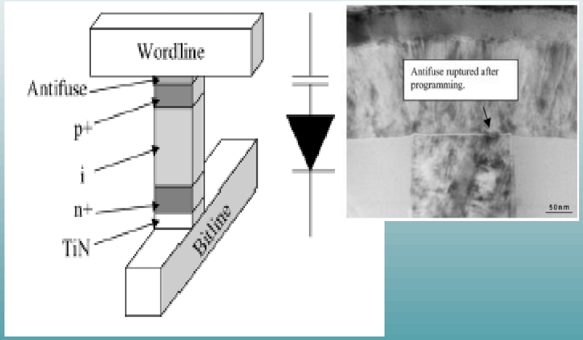


Device Level

- Stanford
- Besang

Matrix: Vertical TFT

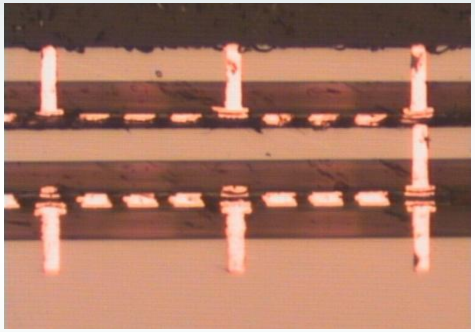


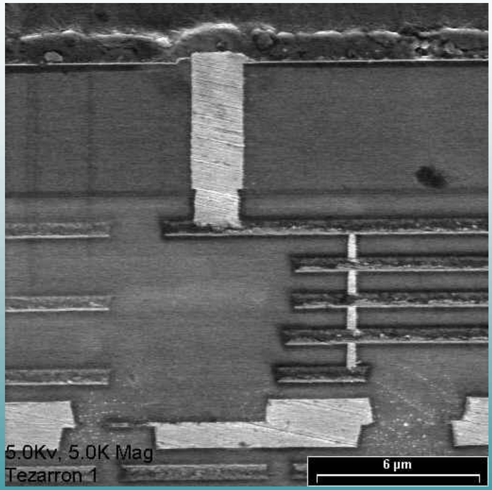


Wafer Level

- Infineon/IBM
- RPI
- ZyCube

Tezzaron

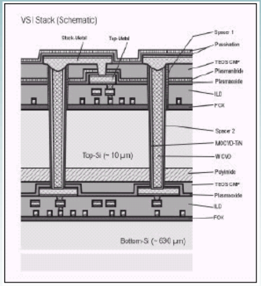


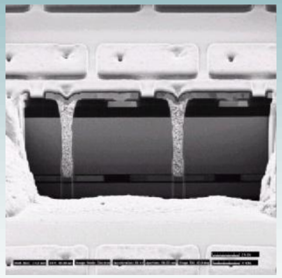


Wafer Level Stacking Approaches

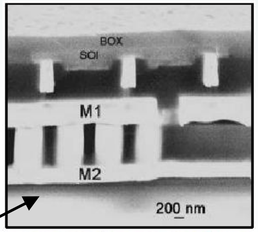
Infineon/IBM

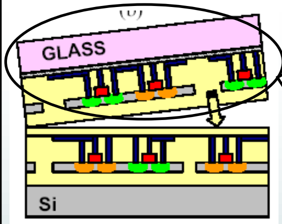
Infineon : W deep via





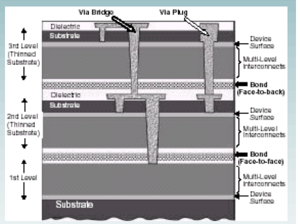
IBM : SOI wafer thinning

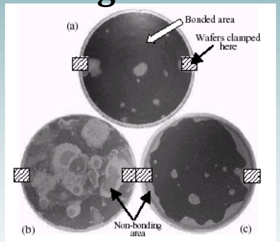




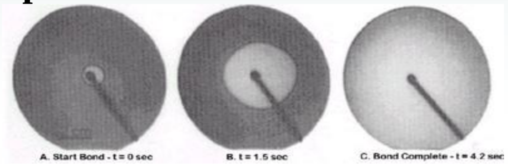
RPI/ Ziptronix/ ZyCube

RPI : Dielectric bonding

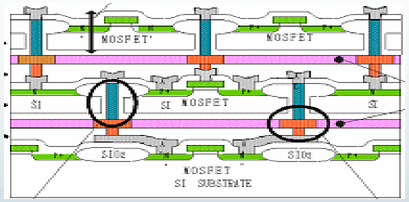




Ziptronix : Covalent bond

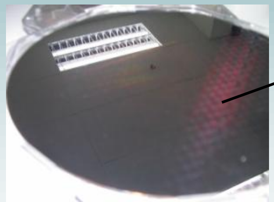


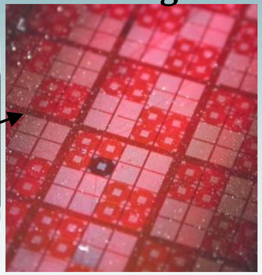
ZyCube : Injection glue bonding



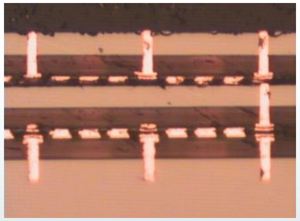
Tezzaron

Tezzaron : Copper bonding

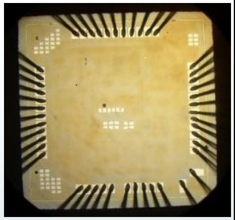




Backside of the stacked wafer



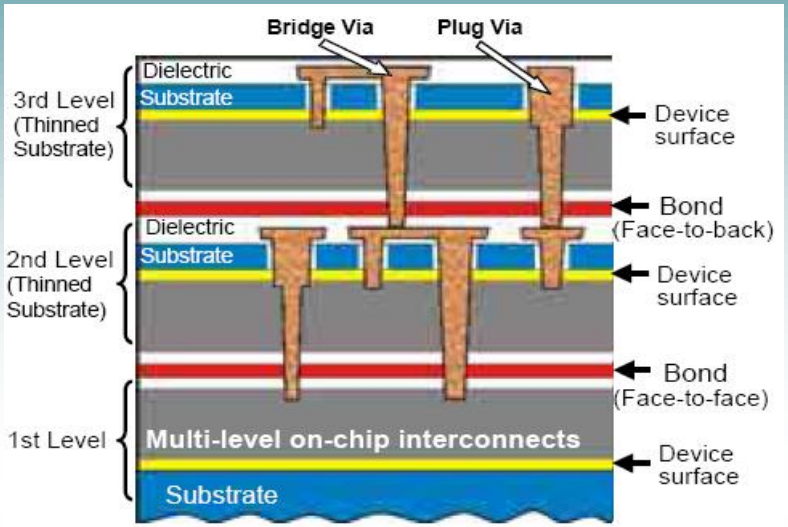
3 wafer stack



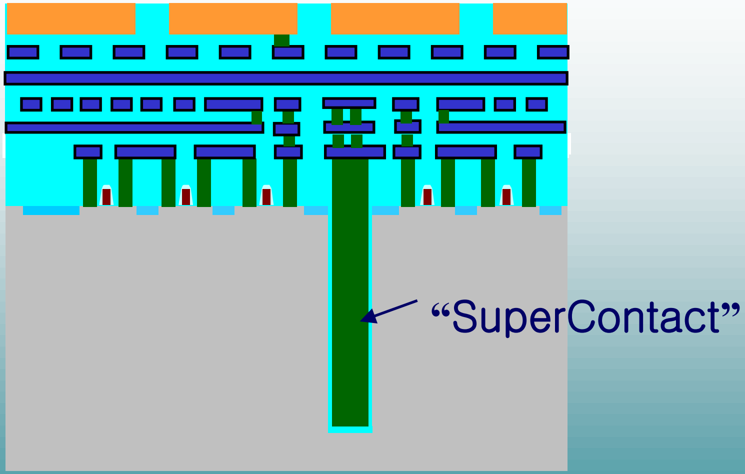
3D Sensor

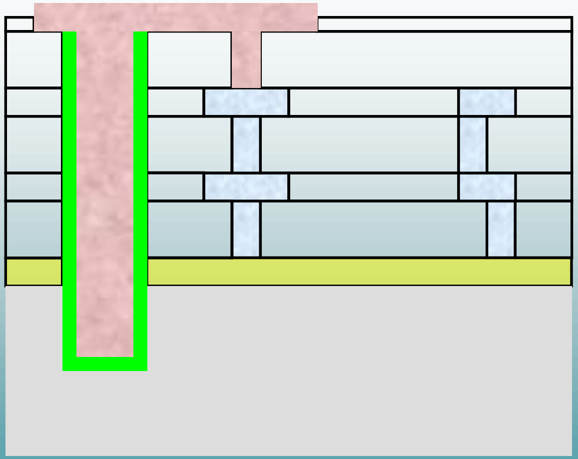
Through-Silicon Via (TSV)

- Via First
- Via Last
- Via at Front end (FEOL)
- Via at Mid line (MOL?)
- Via at Back end (BEOL)



Dr. J.Q. Lu
RPI





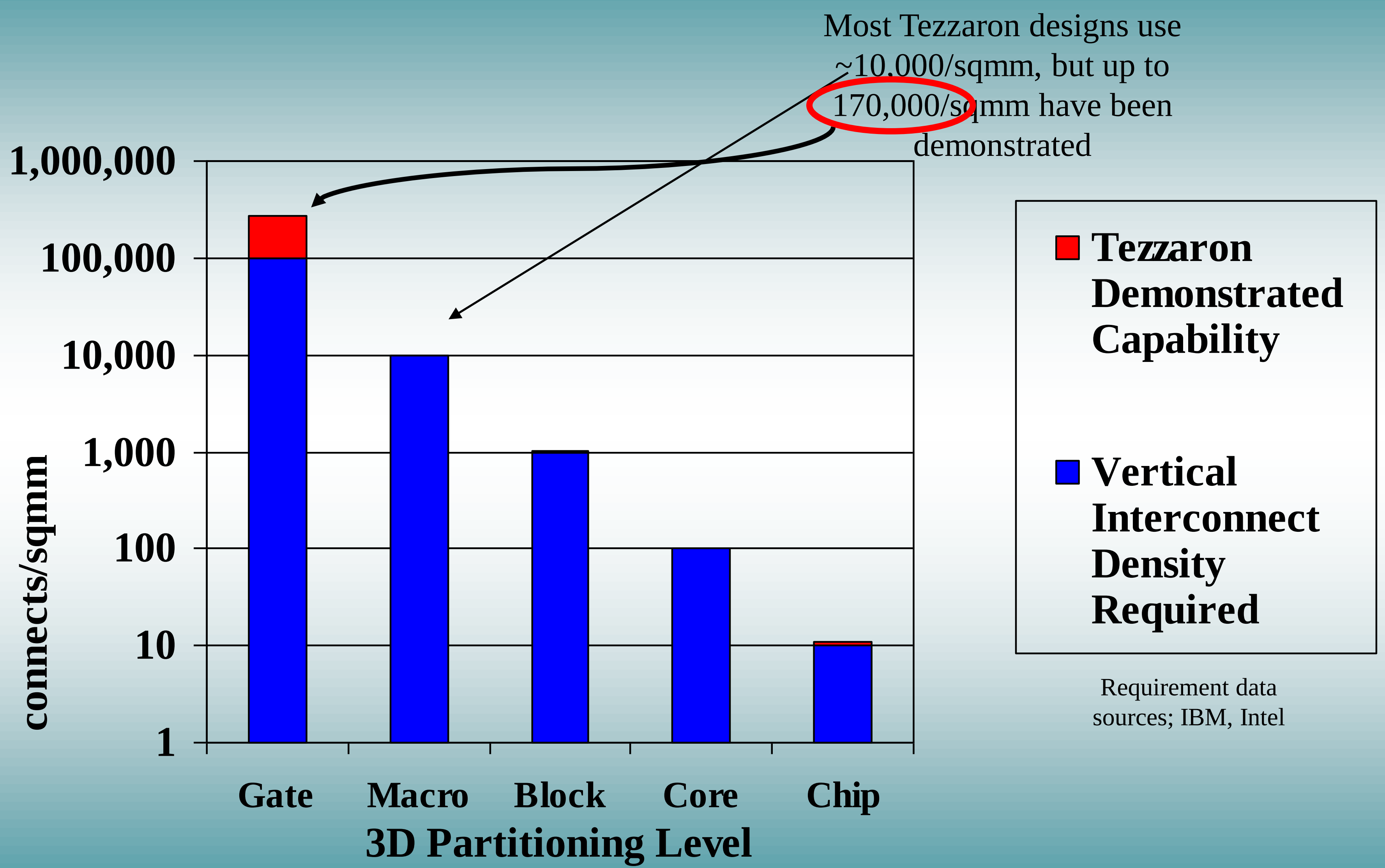
Wafer to Wafer - Best Fit

- Memory
  - DRAM
  - PCRAM, FERAM, MRAM
- FPGA
- Sensors
- Processors
  - Short wires
  - Processor-In-Memory
- 1,000 to 1,000,000 connections per sqmm

Chip to Wafer - Best Fit

- Memory to Logic
- Mixed Materials (GaAs, InP)
- Known Good Die yield++
- 10 to 10,000 connections per sqmm

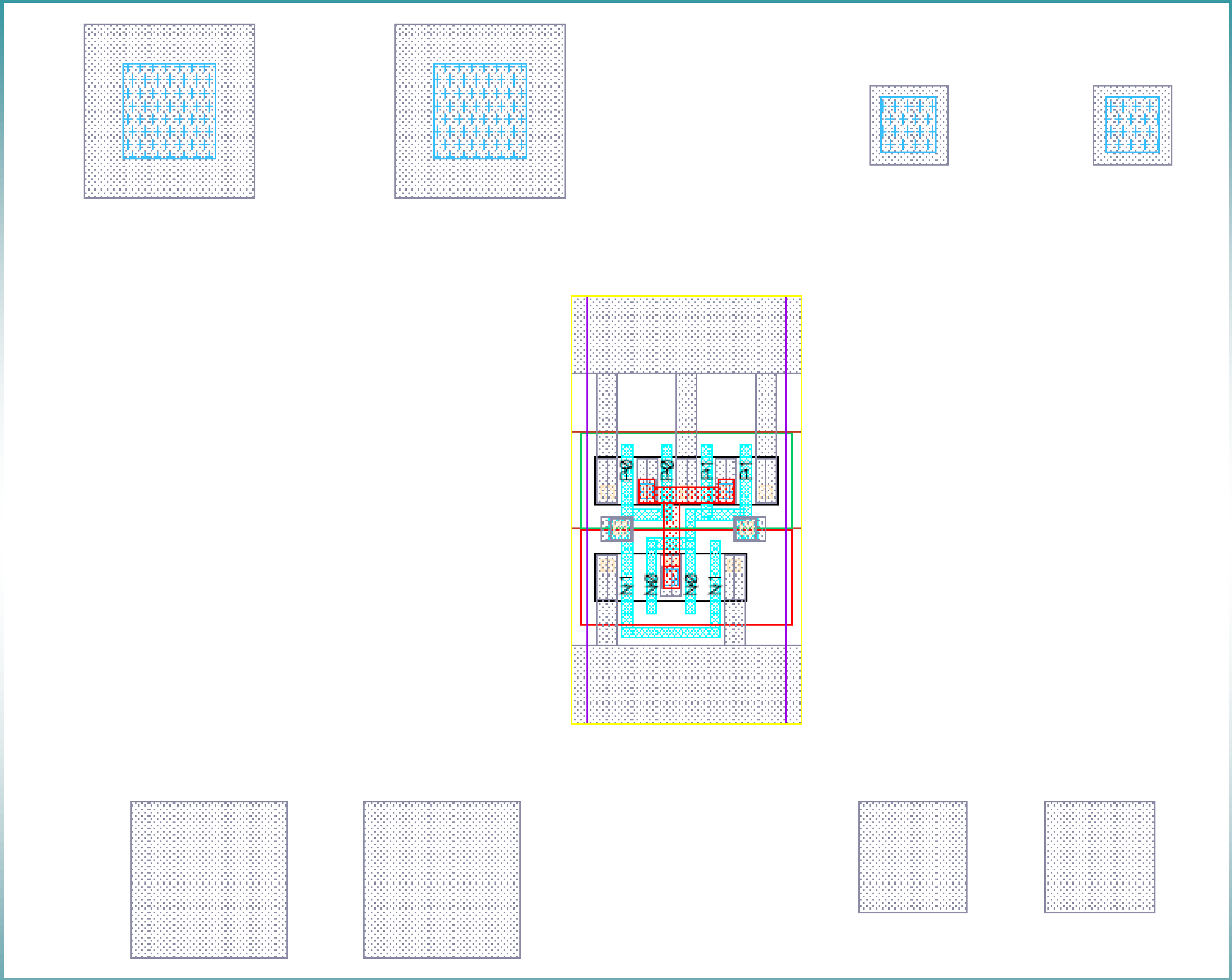
How many interconnects are required?

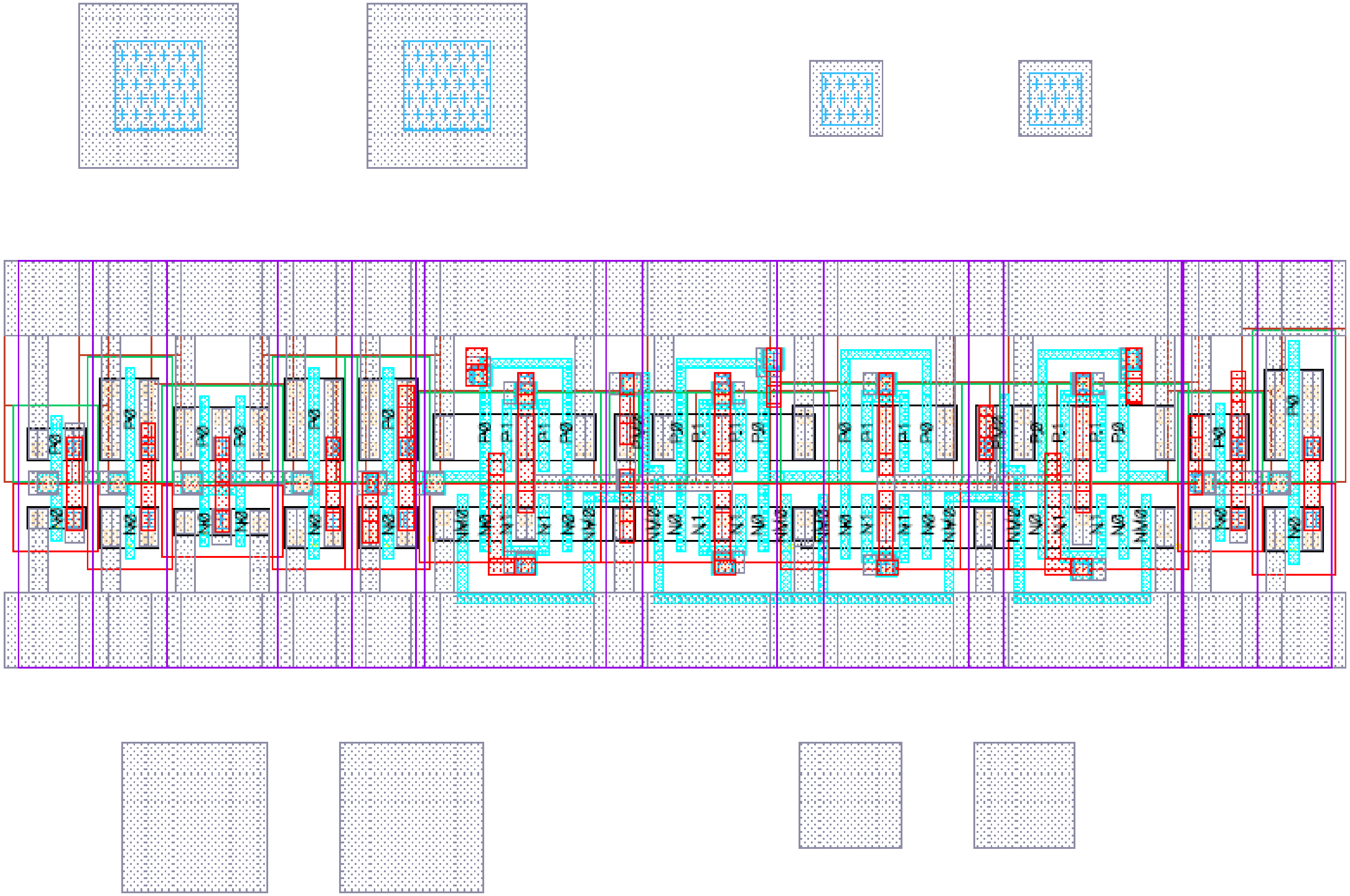


Requirement data sources; IBM, Intel

3D Interconnect Characteristics

	SuperVia™ Via First, BEOL	SuperContact™ 200mm Via First, FEOL	SuperContact™ 300mm Via First, FEOL	Bond Points	Chip to Wafer
Size L X W X D Material	4.0 µ X 4.0 µ X 12.0µ Cu	1.2 µ X 1.2 µ X 6.0µ W	1.6 µ X 1.6 µ X 10.0µ W	1.7 µ X 1.7 µ Cu	10 µ X 10 µ Cu
Minimum Pitch	6.08 µ	<2.5 µ	<3.2 µ	2.4 µ	25 µ
Feedthrough Capacitance	7fF	2-3fF	6fF	<<	<25fF
Series Resistance	<0.25 Ω	<0.6 Ω	<1.5 Ω	<	<





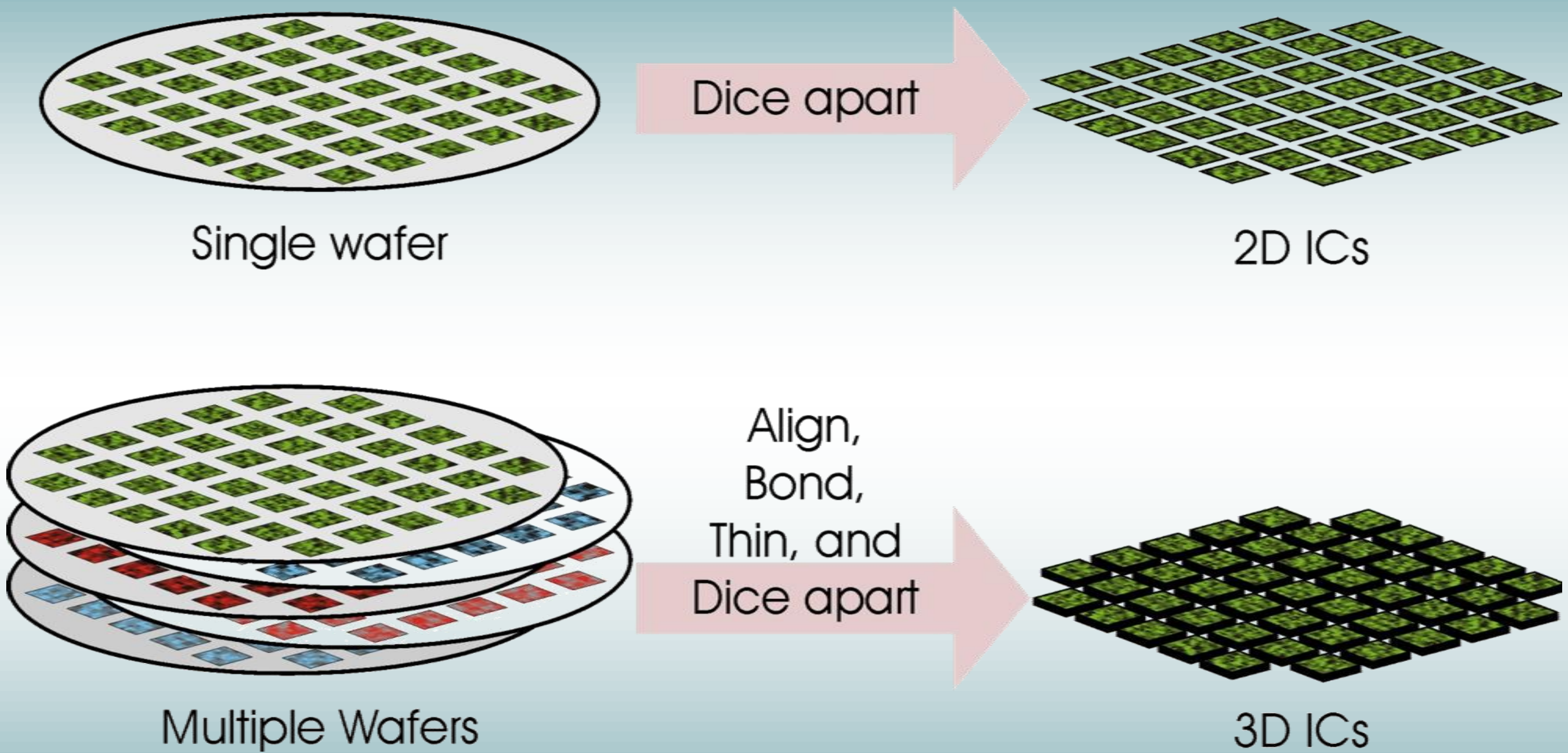
Pitch and Interconnect

- SuperContact™ is $500f^2$ (including spacing)
- Face to face is $350f^2$ (including spacing)
- Chip on wafer I/O pitch is $35{,}000f^2$
- Standard cell gate is 200 to $1000f^2$
  - 3 connections
- Standard cell flip-flop is $5000f^2$
  - 5 connections
- 16 bit sync-counter is $125{,}000f^2$
  - 20 connections
- Opamp is $300{,}000f^2$
  - 4 connections

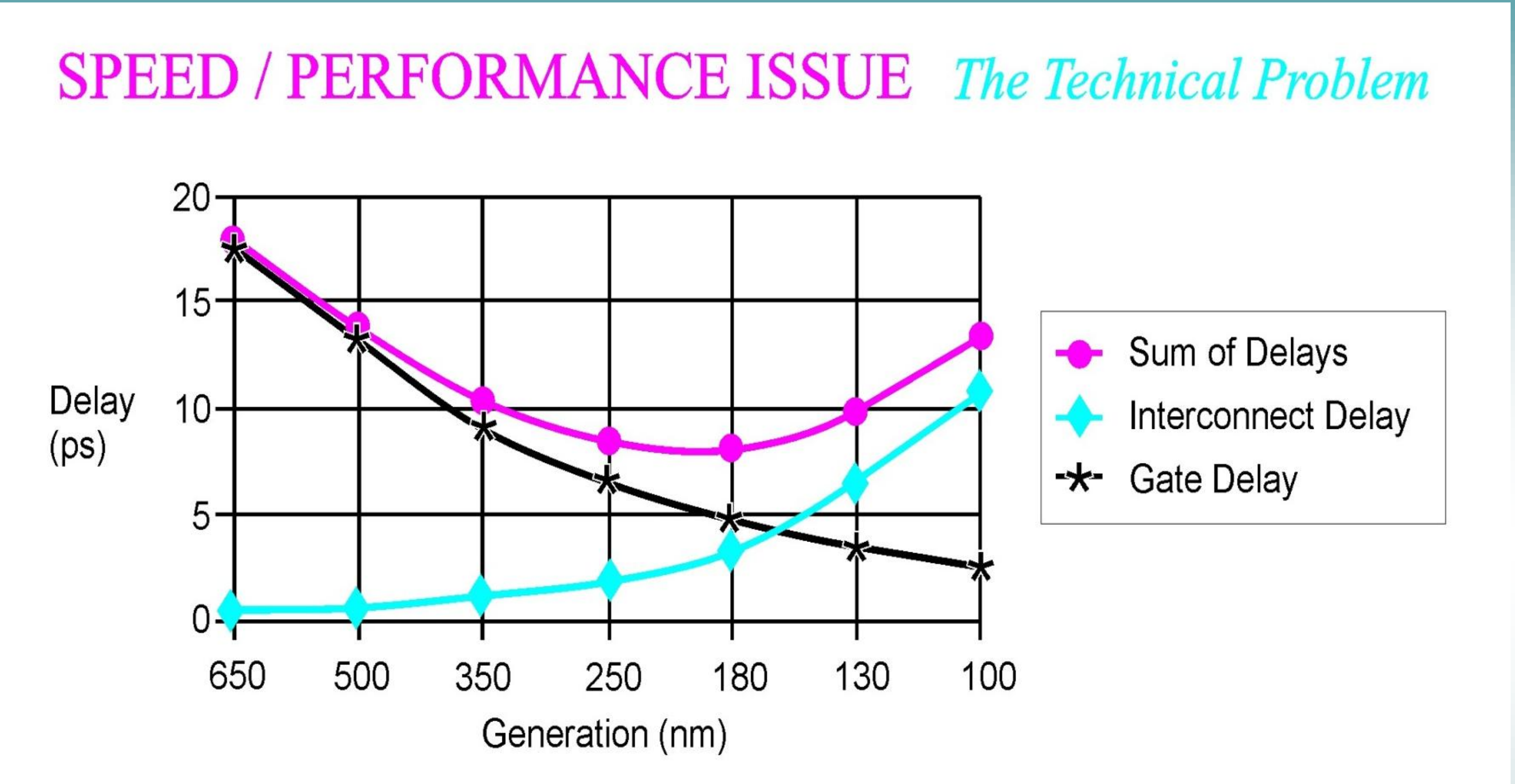
What can 3D achieve?

- Denser
- Faster
- Lower power
- Lower cost
- Higher yield

Denser!



Dice apart
Single wafer
2D ICs
Align,
Bond,
Thin, and
Dice apart
Multiple Wafers
3D ICs



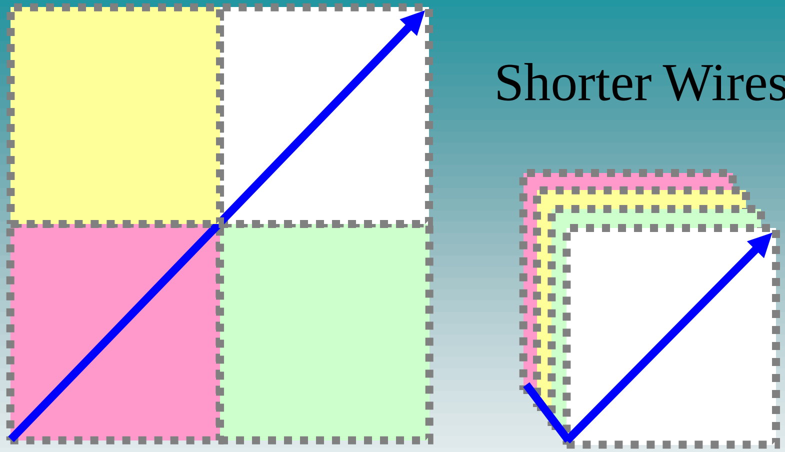
"It is clearly seen in Figure 1, that without further reductions in interconnect delay, reducing gate dimensions much below 130nm do not result in corresponding chip improvements."

NSA Tech Trends Q3 2003

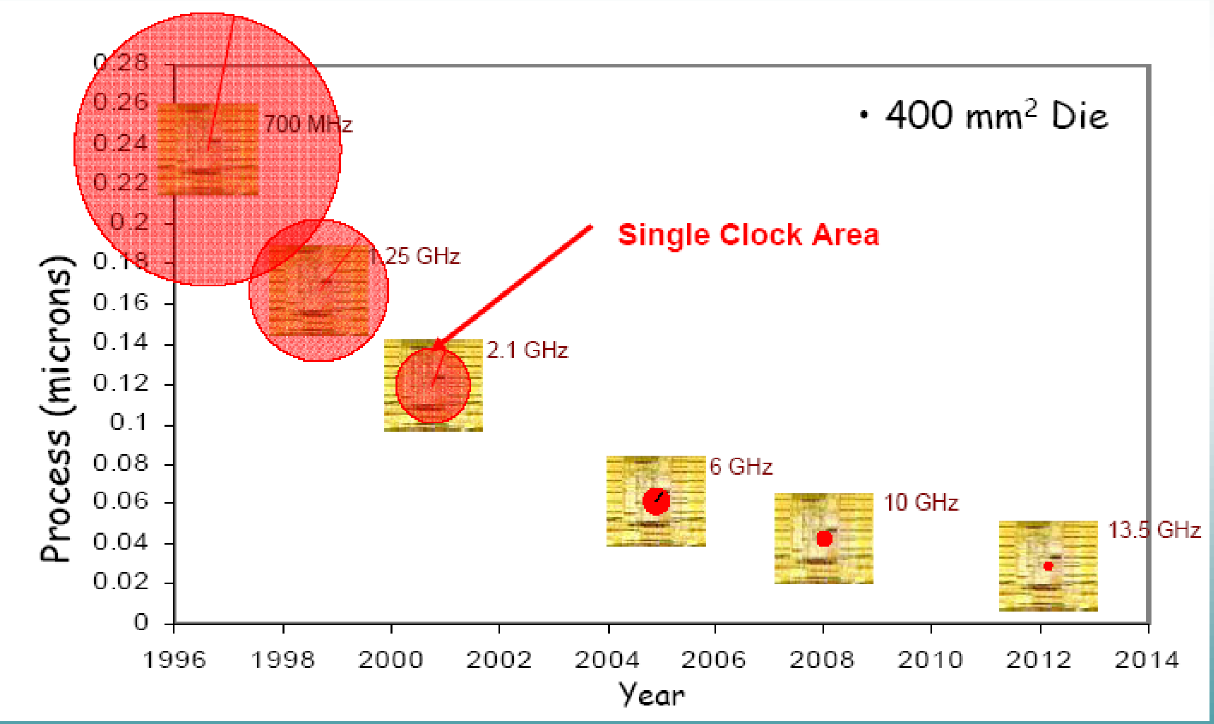
Faster!

Propagation delay is proportional to: $\frac{1}{\text{\# of layers}}$

$$t_d \approx 0.35 \text{ x } rcl^2$$



Shorter Wires



- Global Interconnect "problem"
  - Span of Control

Lower Power!

$$P_{avg} = VDD \text{ x } I_{avg} = C_{tot} \text{ x } VDD^2 \text{ x } f_{clk}$$

C is mostly due to wiring

Therefore:

$$P_{avg} \propto l_{avg}$$

Or:

$$P_{avg\ stacked} \approx \frac{P_{avg\ single\ layer}}{\#\text{ of layers}}$$

++ *Reduction of repeaters*

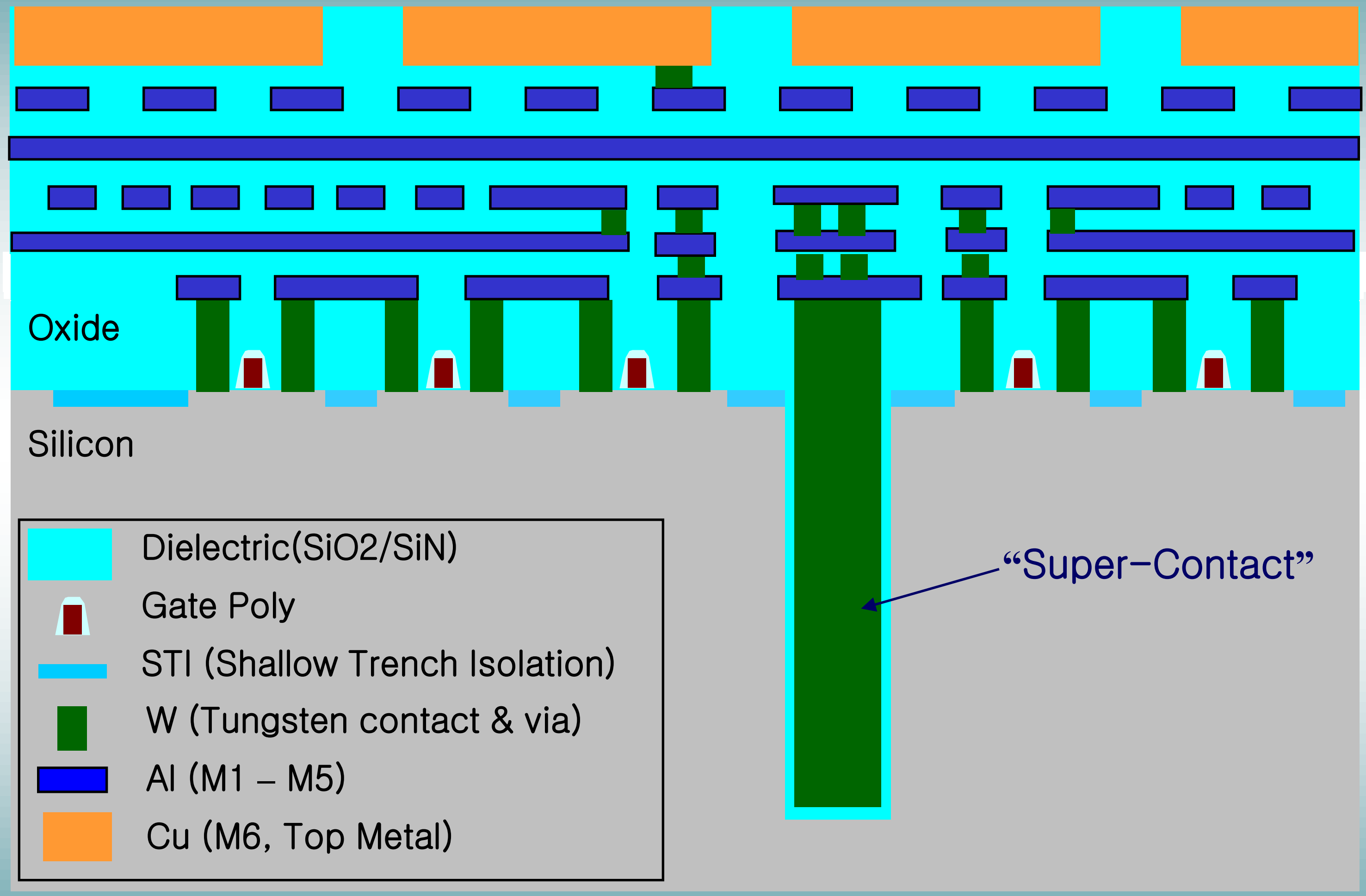
Operation	Energy
32-bit ALU operation	5 pJ
32-bit register read	10 pJ
Read 32 bits from 8K RAM	50 pJ
Move 32 bits across 10mm chip	100 pJ
Move 32 bits off chip	1300 to 1900 pJ

Calculations using a 130nm process operating at a core voltage of 1.2V
(Source: Bill Dally, Stanford)

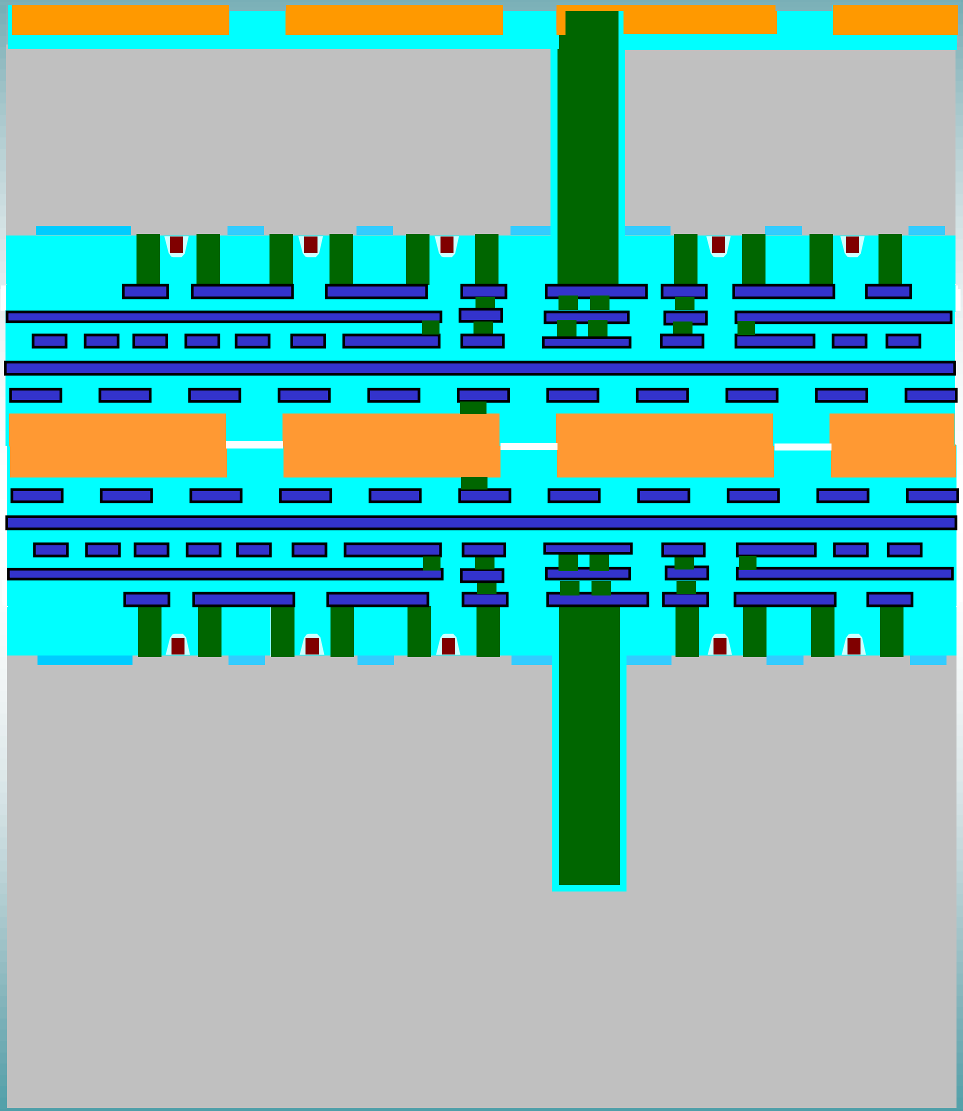
Lower Costs & Higher Yield!

- Better optimization per wafer
- Less processing per layer
- Higher bit density in memories
- Lower test cost using Bi-STAR™
- Higher yield using Bi-STAR™

A Closer Look at Wafer-Level Stacking



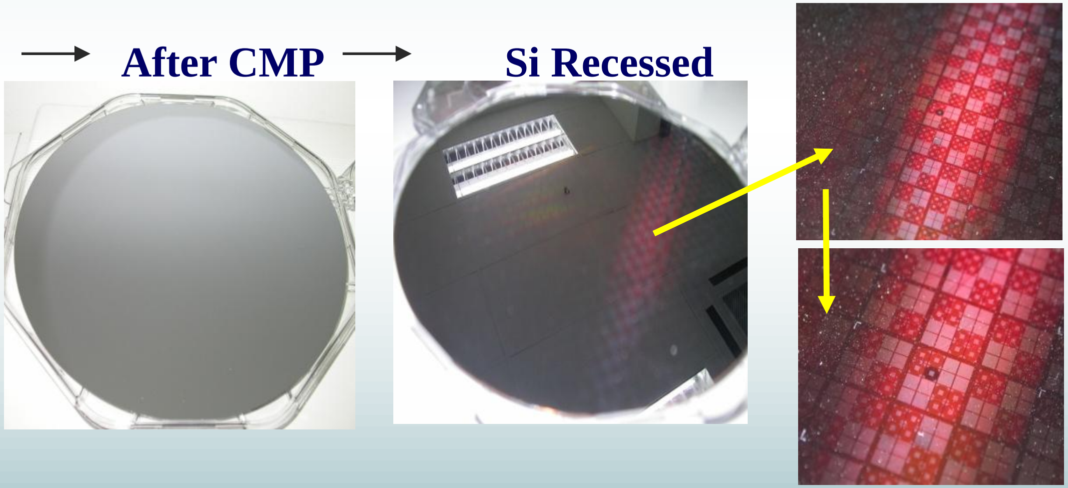
Next, Stack a Second Wafer & Thin:



Stacking Process Sequential Picture

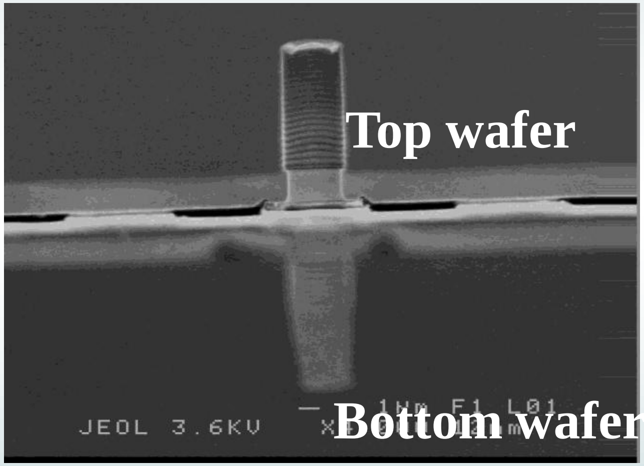
Two wafer Align & Bond → Course Grinded → Fine Grinded

→ After CMP → Si Recessed



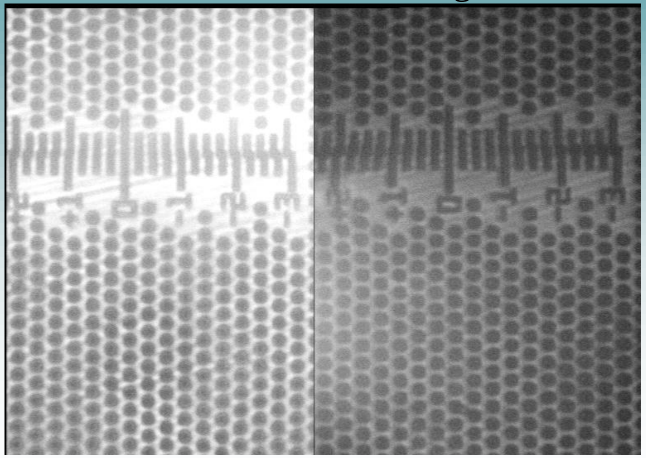
High Precision Alignment

Misalign=0.3um



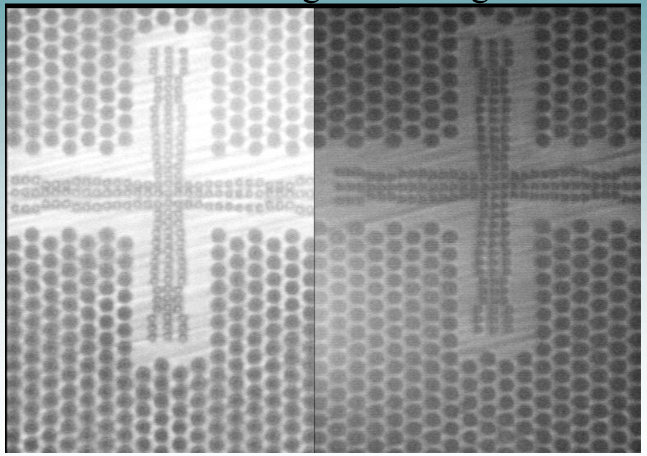
Stacking Alignment, Infra Red Microscope Images

VERNIER_X-axis Misalignment



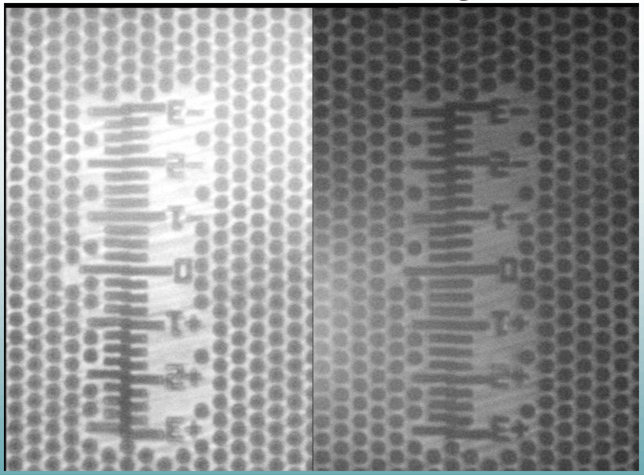
Wafer Left (-80mm,0) Wafer Right (+80mm,0)

Tezzaron's Alignment Target

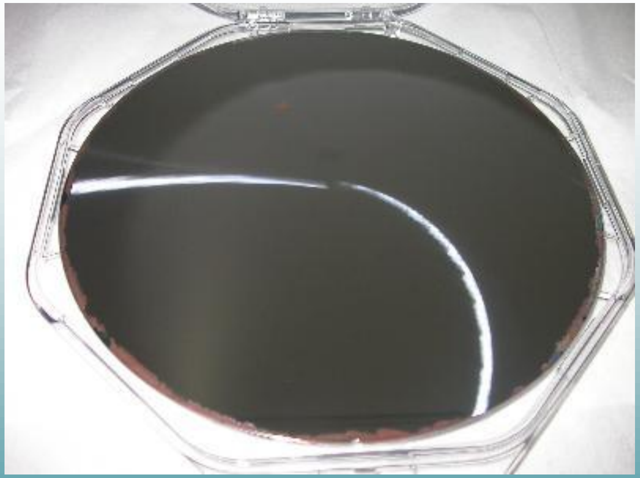


Wafer Left (-80mm,0) Wafer Right (+80mm,0)

VERNIER_Y-axis Misalignment

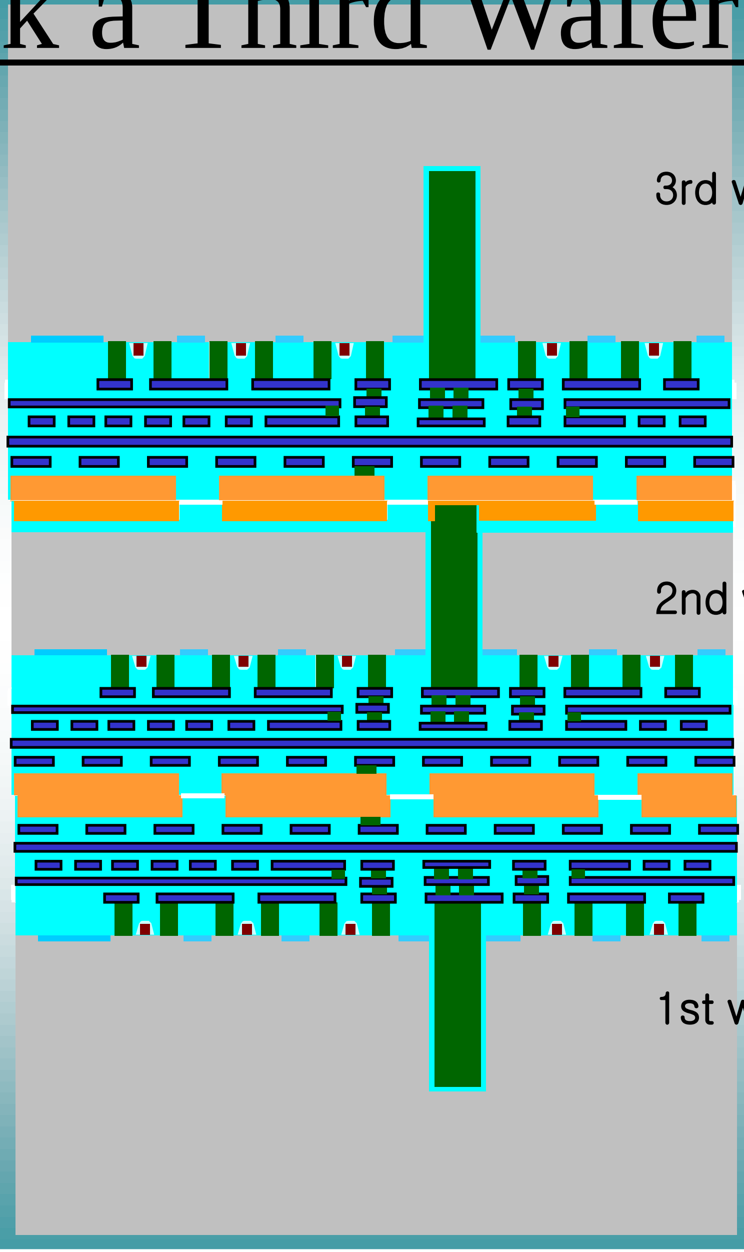


Wafer Left (-80mm,0) Wafer Right (+80mm,0)

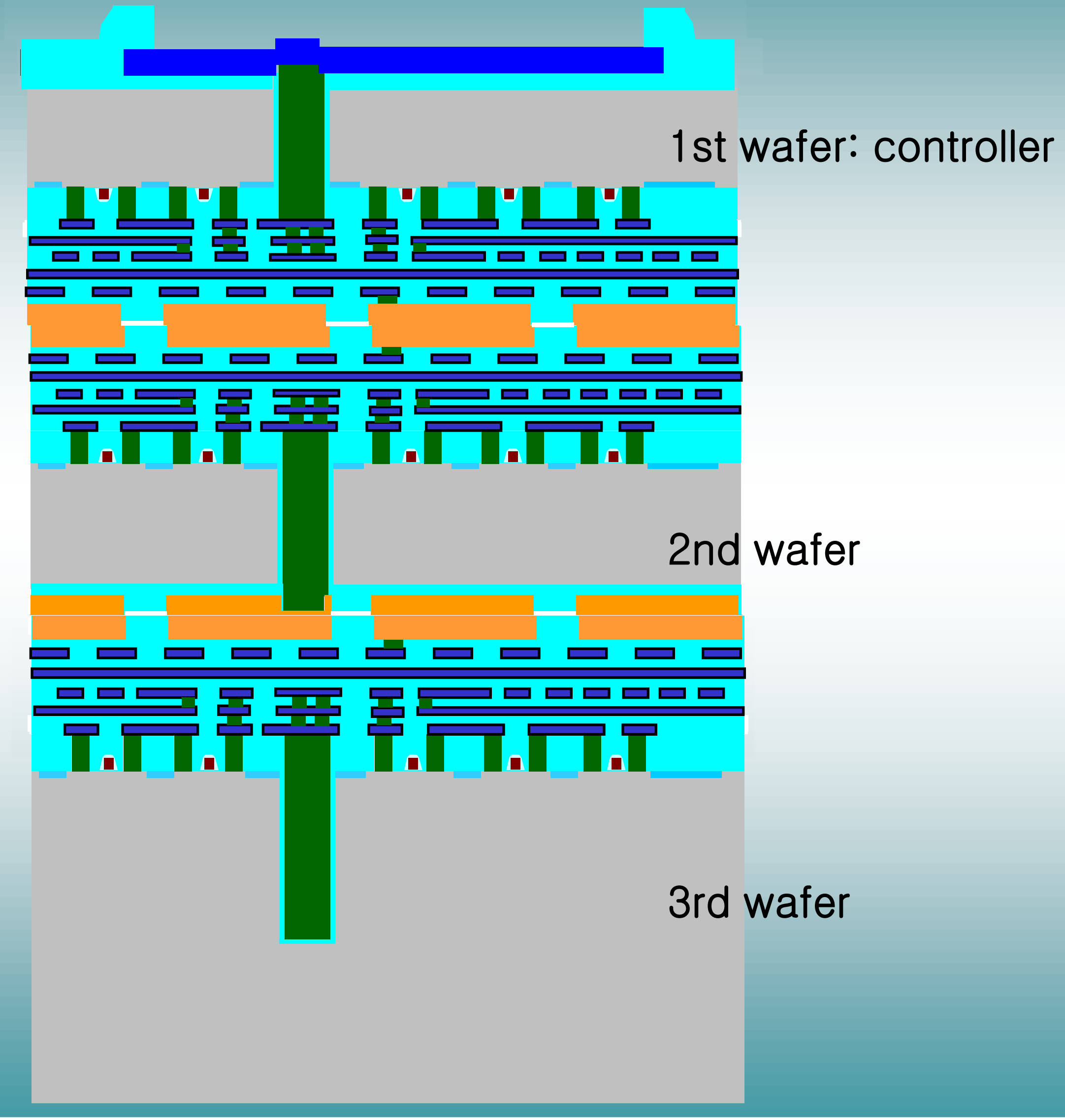


Staked wafer picture

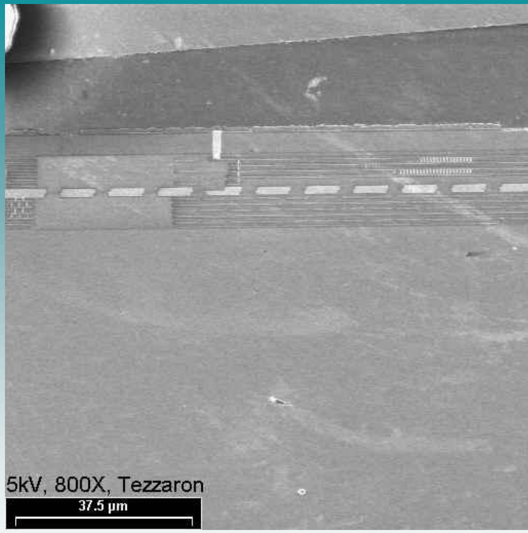
Then, Stack a Third Wafer:

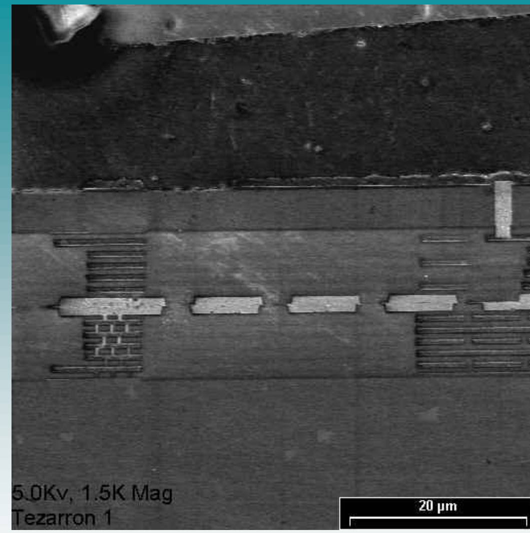


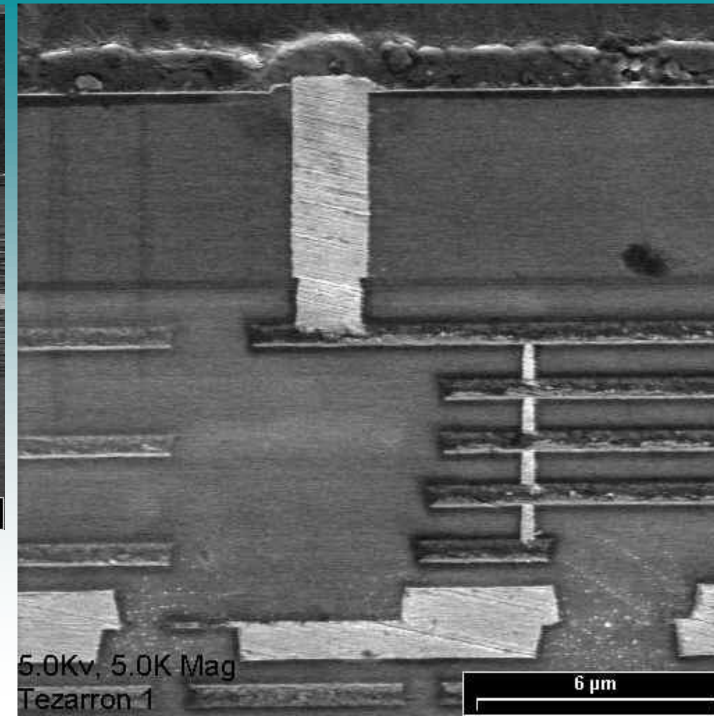
Finally, Flip, Thin & Pad Out:

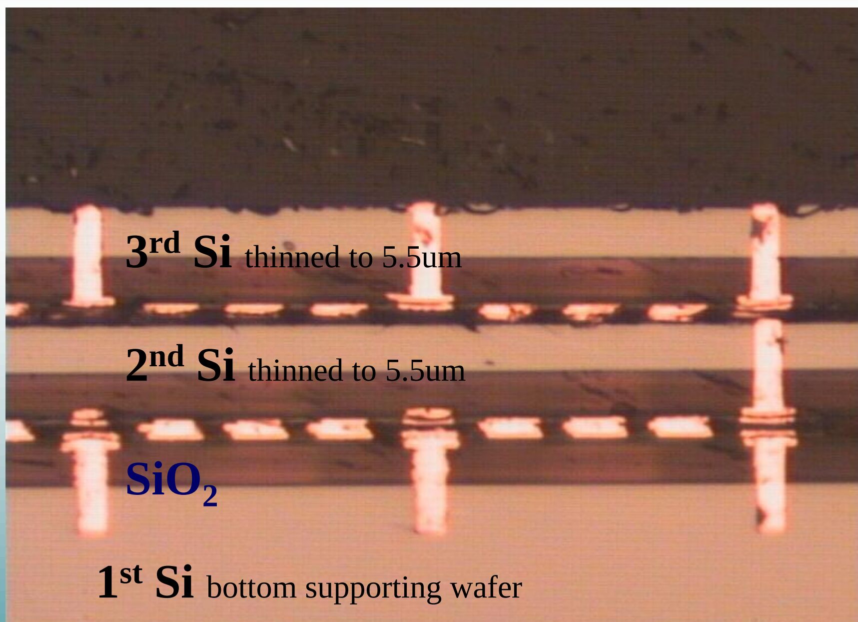


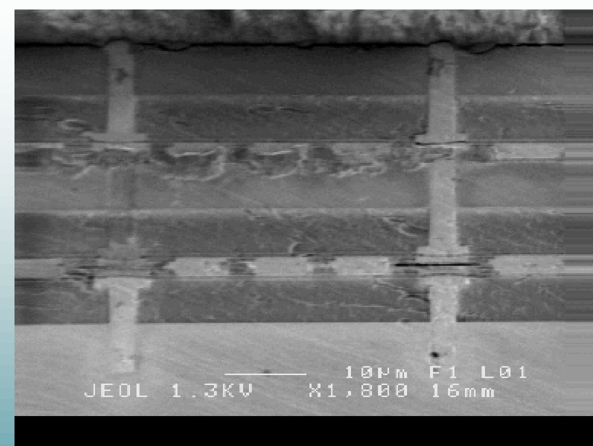
This is the completed stack!


5kV, 800X, Tezzaron
37.5 µm

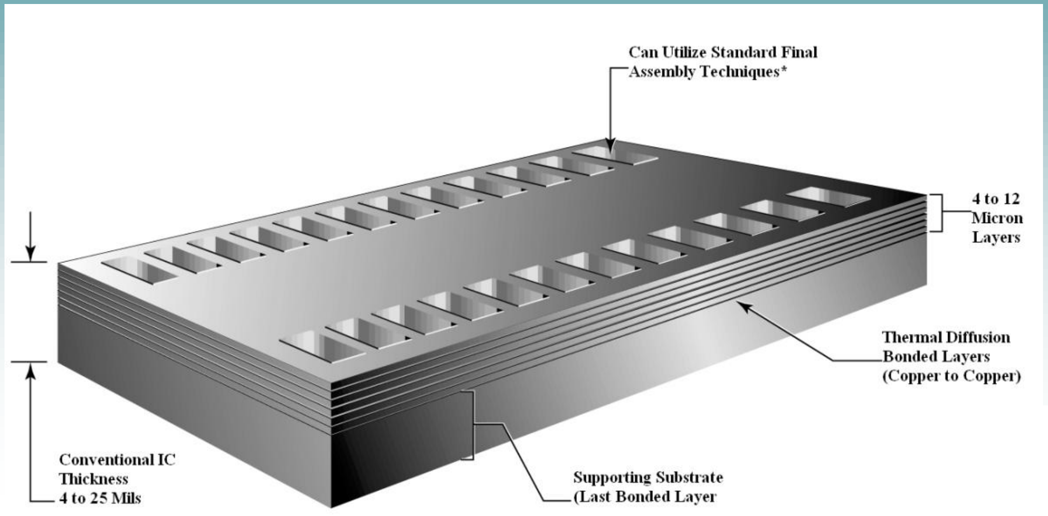

5.0Kv, 1.5K Mag
Tezarron 1
20 µm

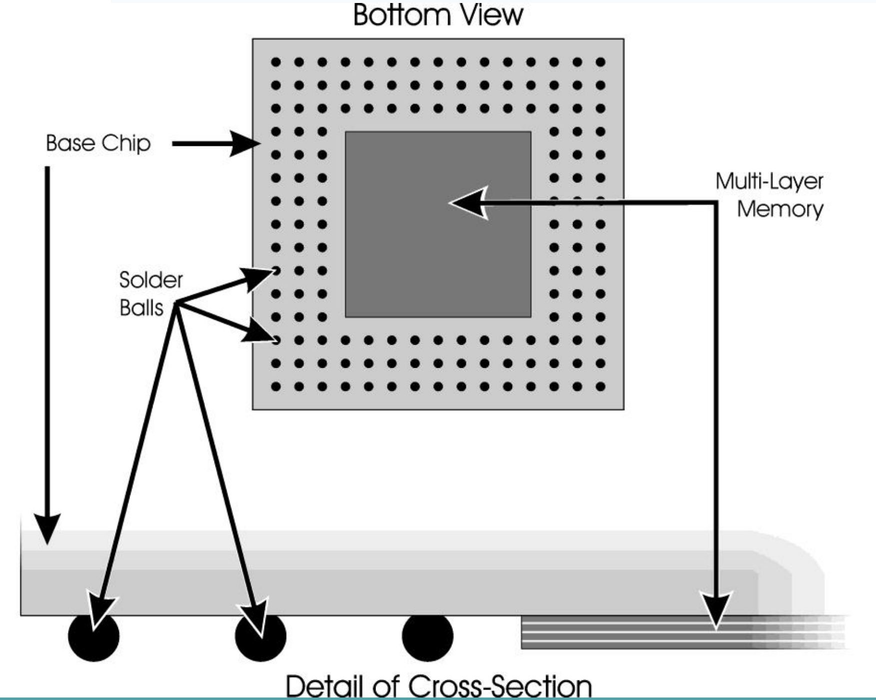

5.0Kv, 5.0K Mag
Tezarron 1
6 µm


3rd Si thinned to 5.5um
2nd Si thinned to 5.5um
SiO2
1st Si bottom supporting wafer

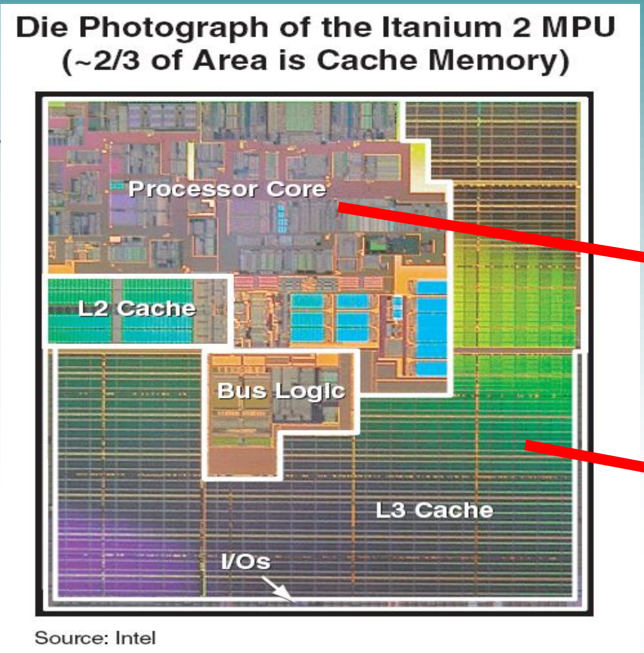

10µm F1 L01
JEOL 1.3KV X1,800 16mm

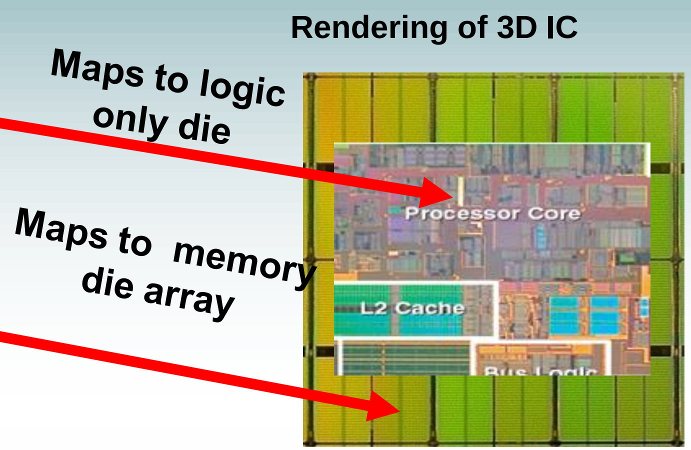
The Next Step





3D Heterogeneous Integration





BEFORE Intel Photo used as proxy

Only Memory Directly Compatible with Logic (virtually no choice!)

AFTER: 3D IC

14x increase in memory density

4X Logic Cost Reduction

29x → 100x memory cost reduction (choice!)

Single Die~ 430 mm2 2D IC "All or Nothing"

Wafer Cost ~ $6,000

Low yield ~ 15%, ~ 10 parts per wafer

memory costs ~ $44/MB

128MB not 9MB

memory costs ~ $1.50/MB → $0.44/MB

Memory on Die Power Advantage

Operation	Energy
32-bit ALU operation	5 pJ
32-bit register read	10 pJ
Read 32 bits from 8K RAM	50 pJ
Move 32 bits across 10mm chip	100 pJ
Move 32 bits off chip	1300 to 1900 pJ

Calculations using a 130nm process operating at a core voltage of 1.2V
(Source: Bill Dally, Stanford)

DDR3 ~40mW per pin

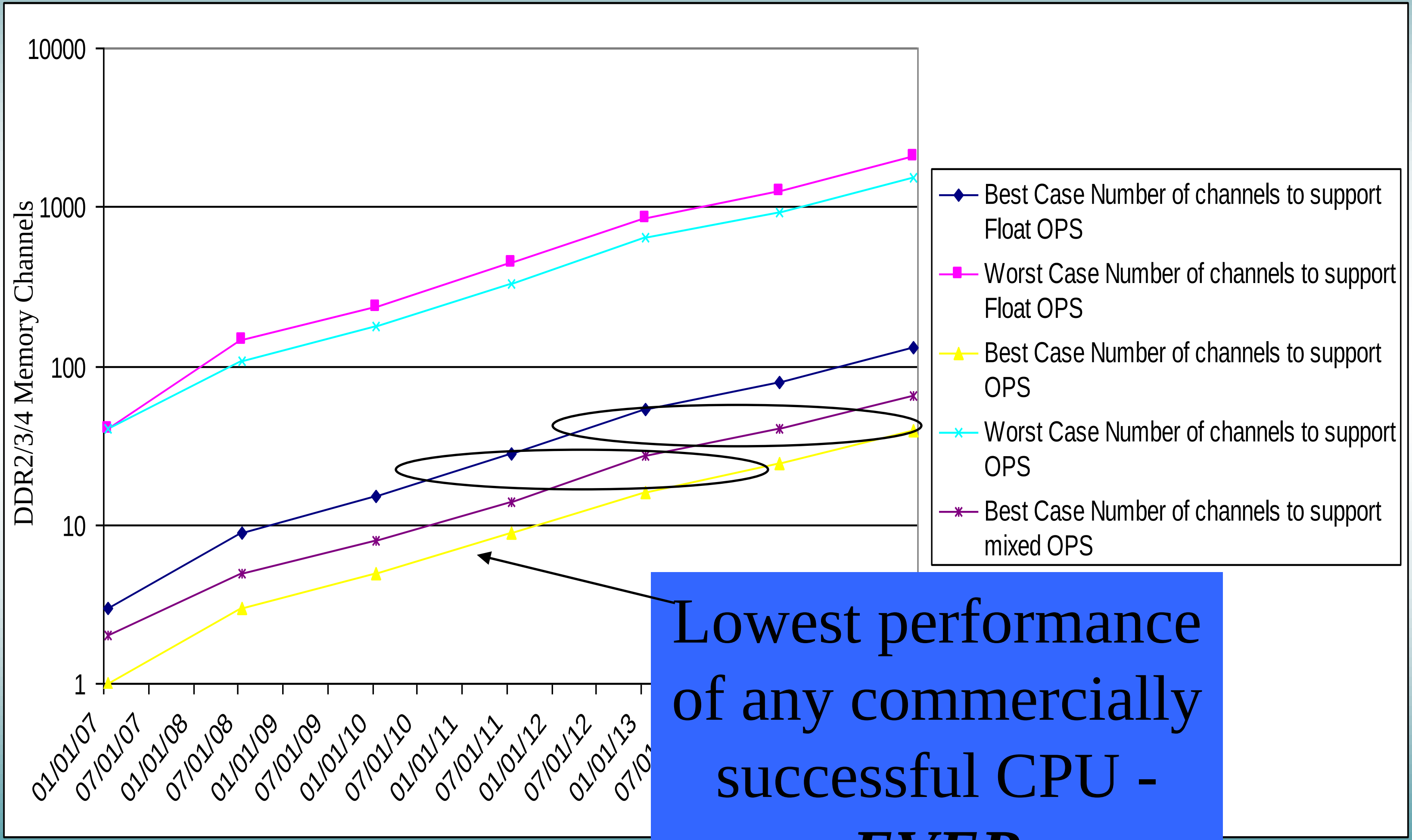
1024 Data pins →40W

4096 Data pins →160W

Die on Wafer ~24uW per pin

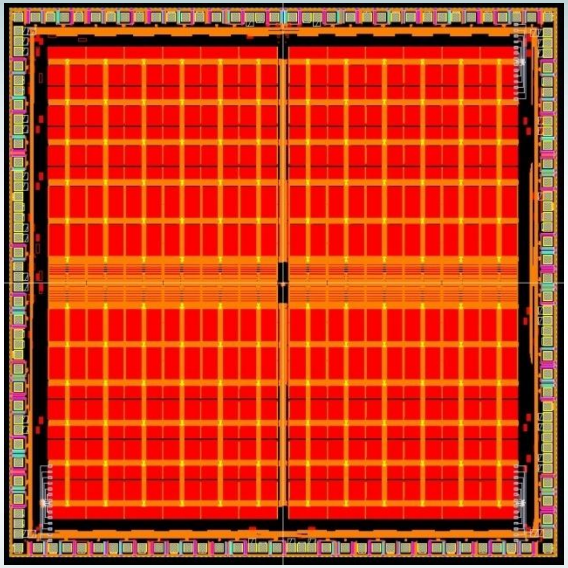
The Bandwidth Crisis:

You know you have a problem when there is a log scale….



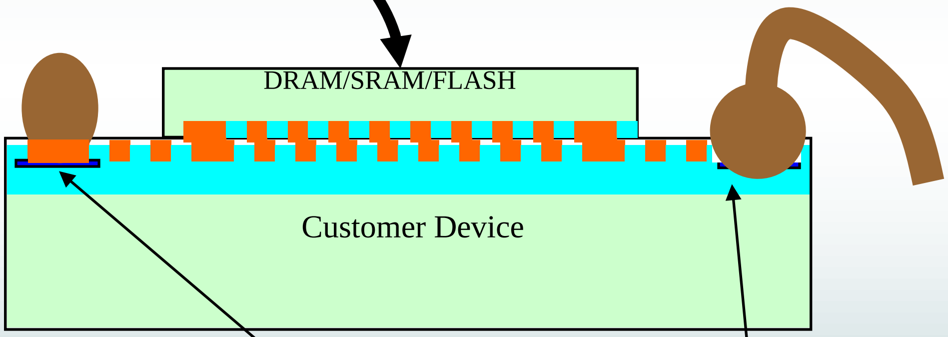
Lowest performance of any commercially successful CPU - ***EVER***

The Killer Application

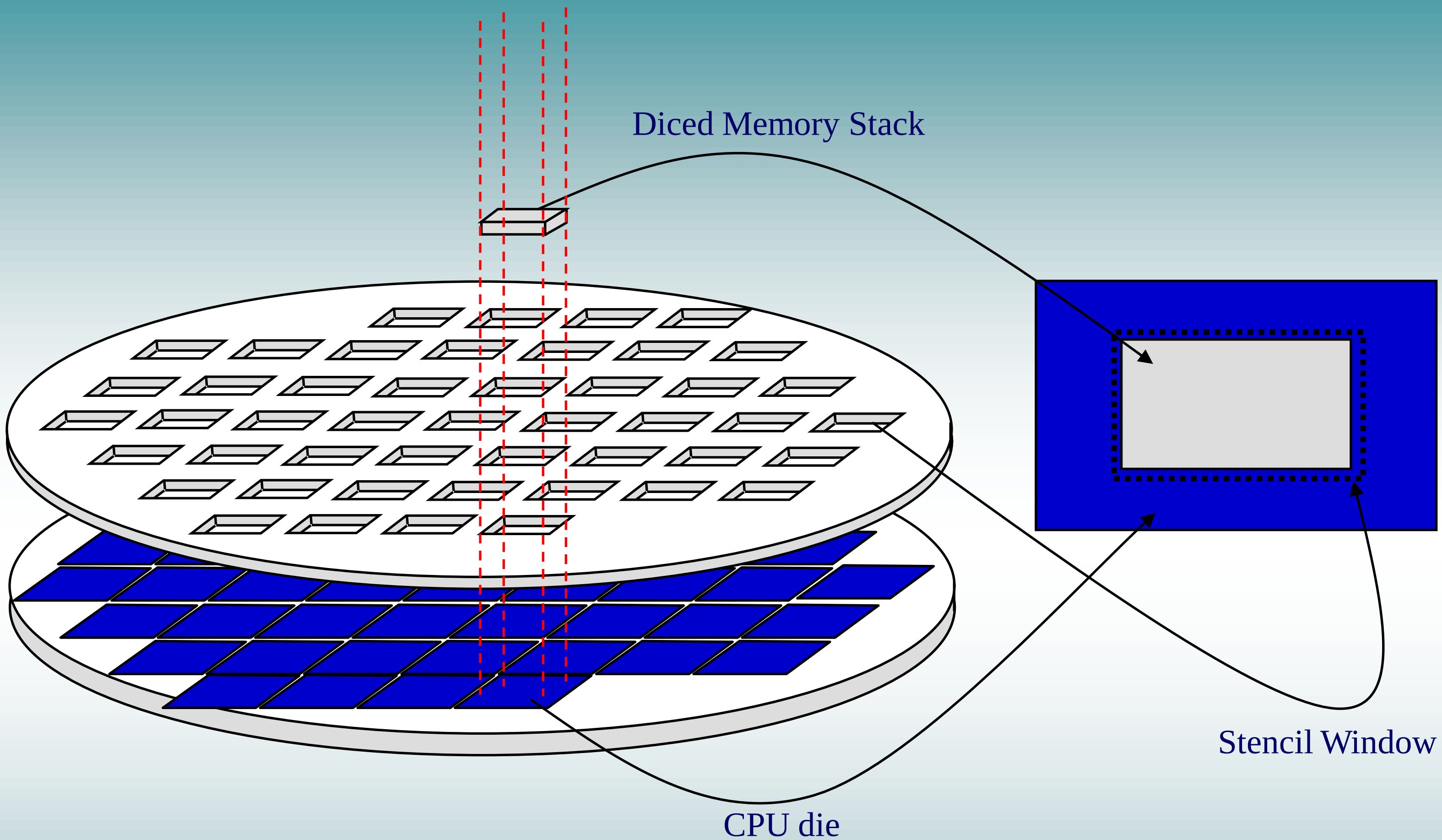


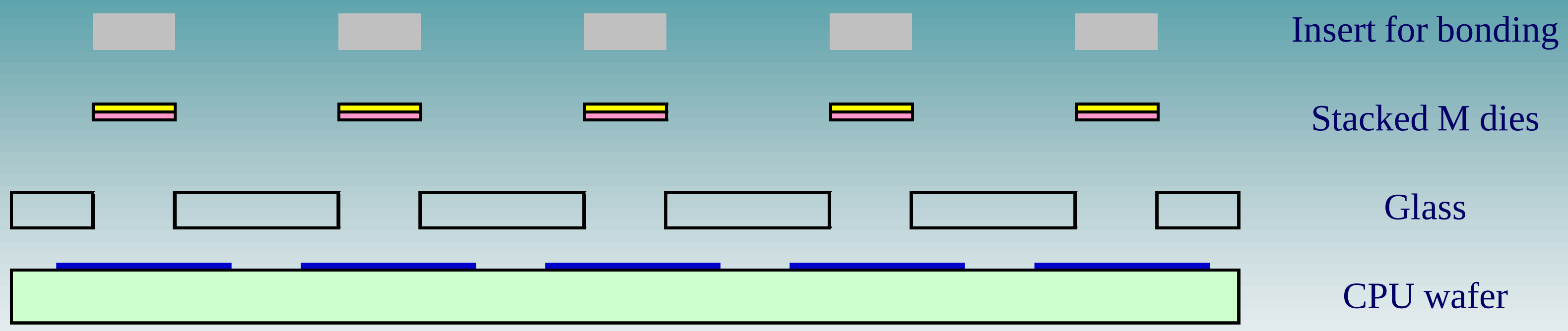
2D or 3D
Memory Device

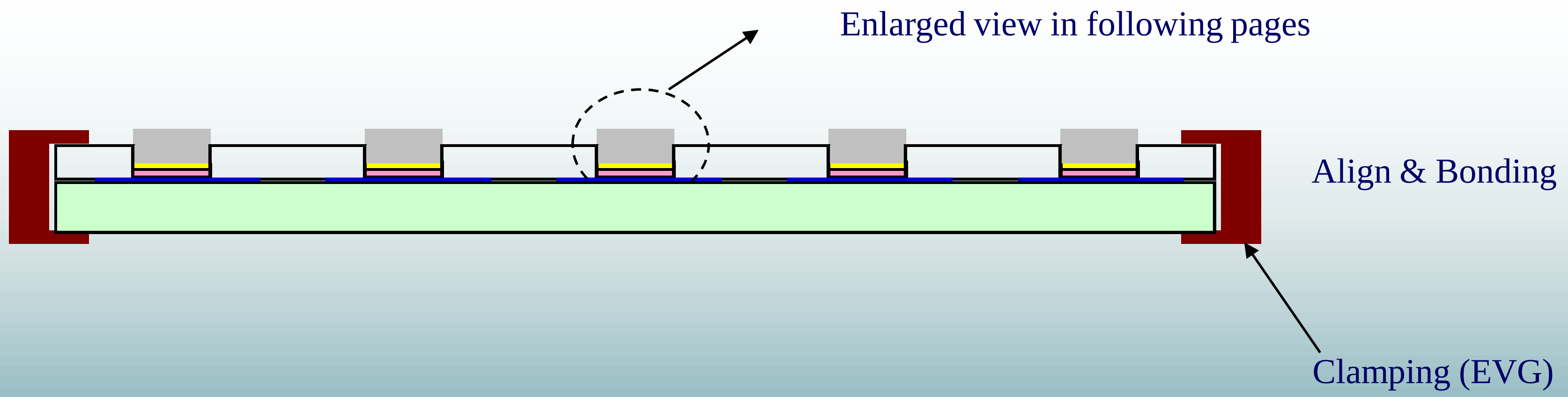
Embedded Performance with
far superior cost/density



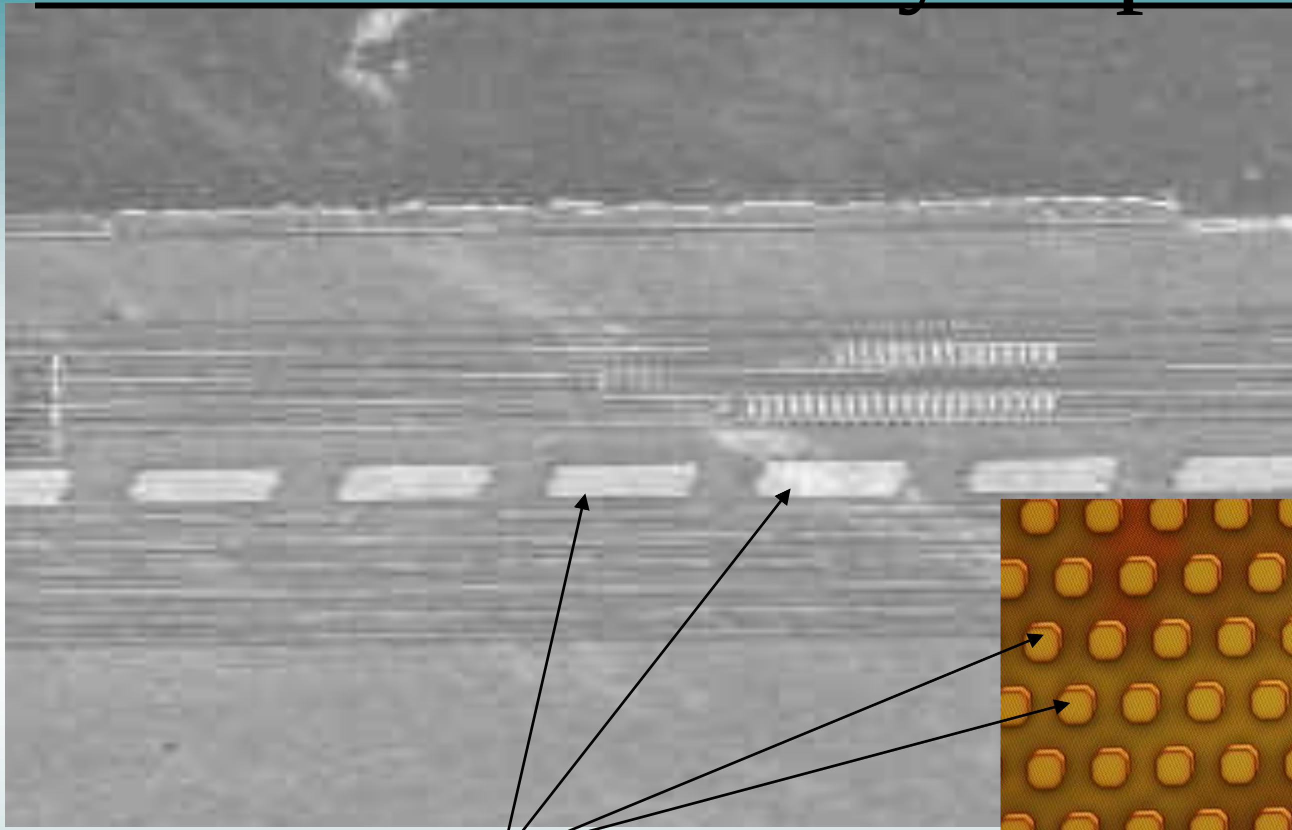
I/O Pad area : Bumping or wire bonding

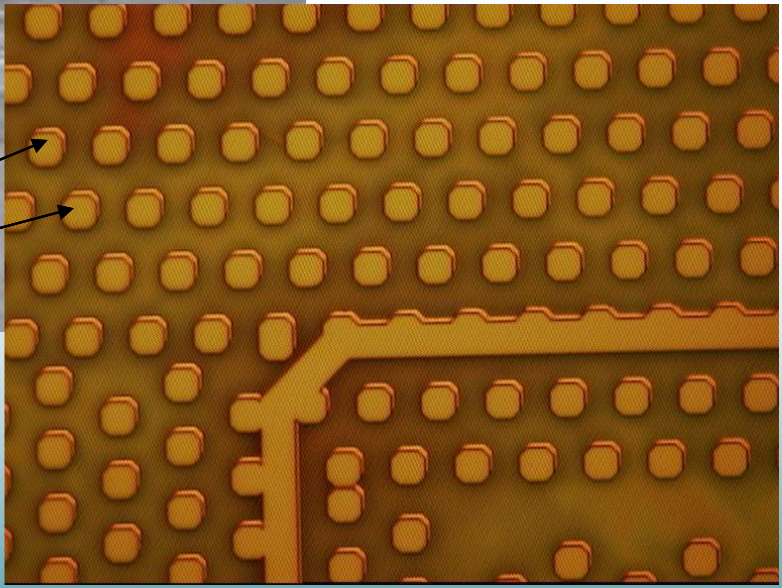

Diced Memory Stack
Stencil Window
CPU die


Insert for bonding
Stacked M dies
Glass
CPU wafer


Enlarged view in following pages
Align & Bonding
Clamping (EVG)

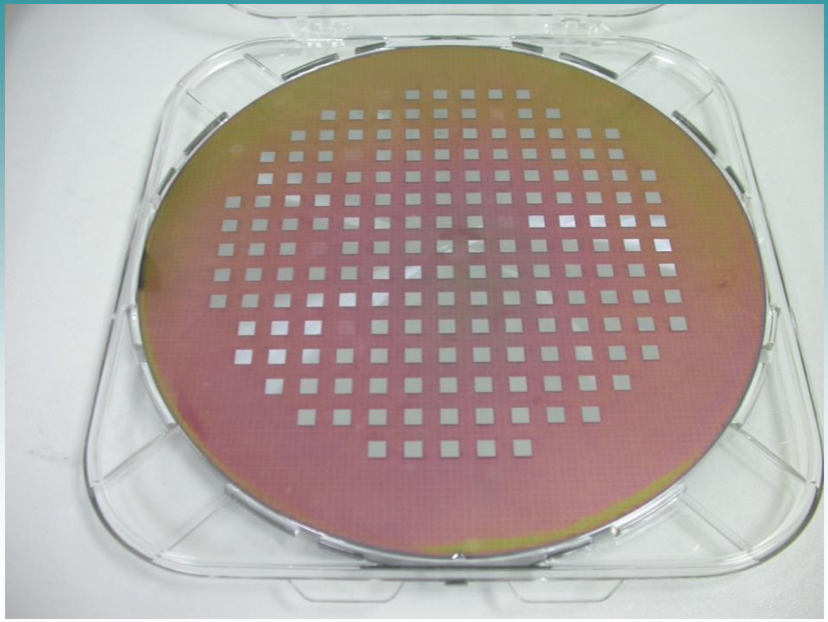
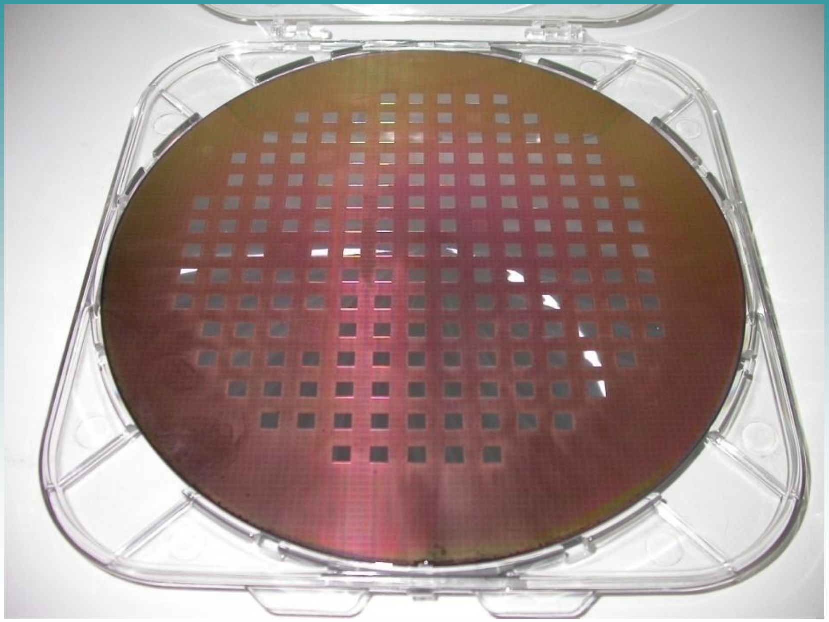
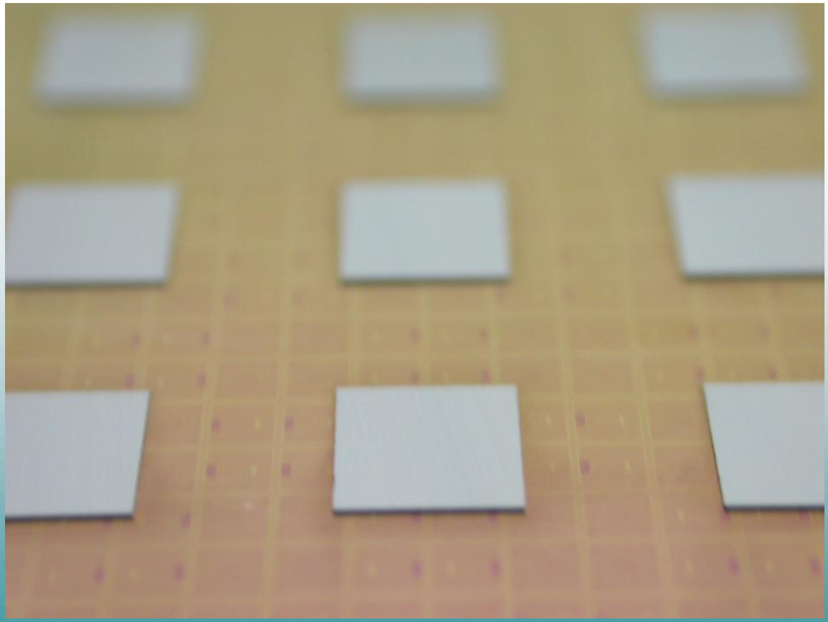
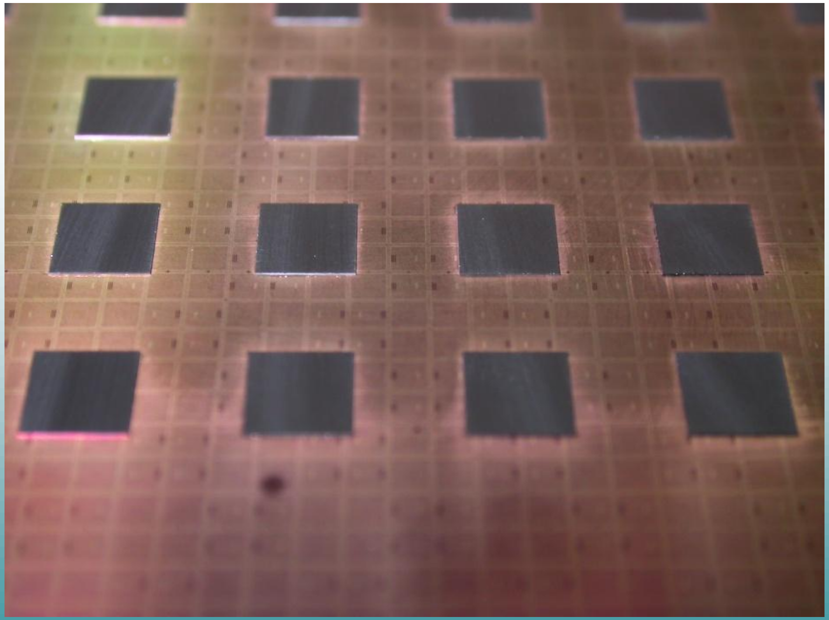
Tezzaron Assembly Capabilities





Interconnect at 10um pitch

10,000 I/O per sqmm

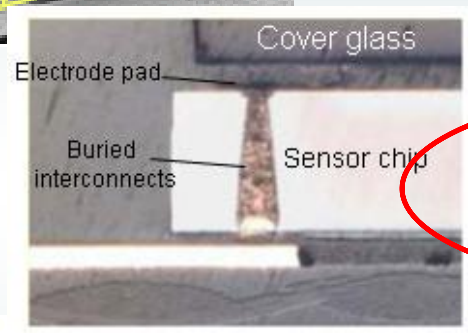





How Real is 3D???

Samsung
16Gb NAND flash (2Gx8 chips), 560µ thin

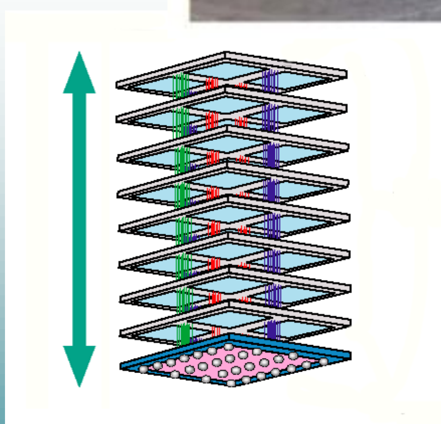
Micron
Osmium Memory Stacking

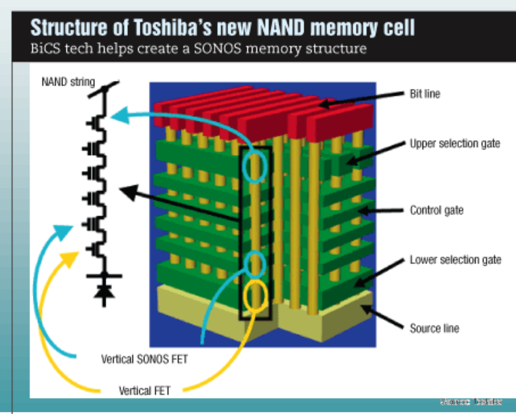
Intel
CPU + memory



OKI/Zycube
CMOS Sensor

NEDO
1Gbit DRAM (128Mbx8 chips) 5mm2





Raytheon/Ziptronix
PIN Detector Device

IBM
RF Silicon Circuit Board / TSV

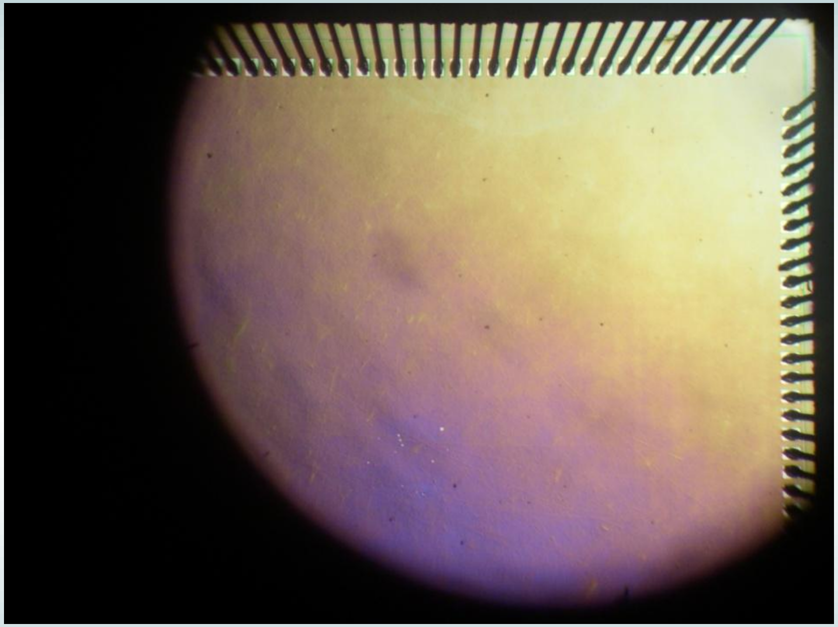
Toshiba
3D NAND

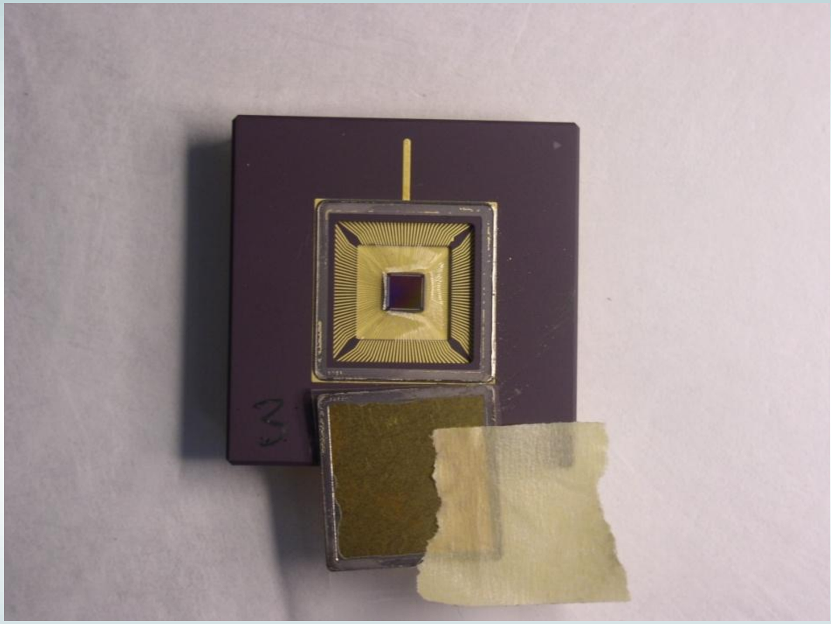
CPU/Memory Stack

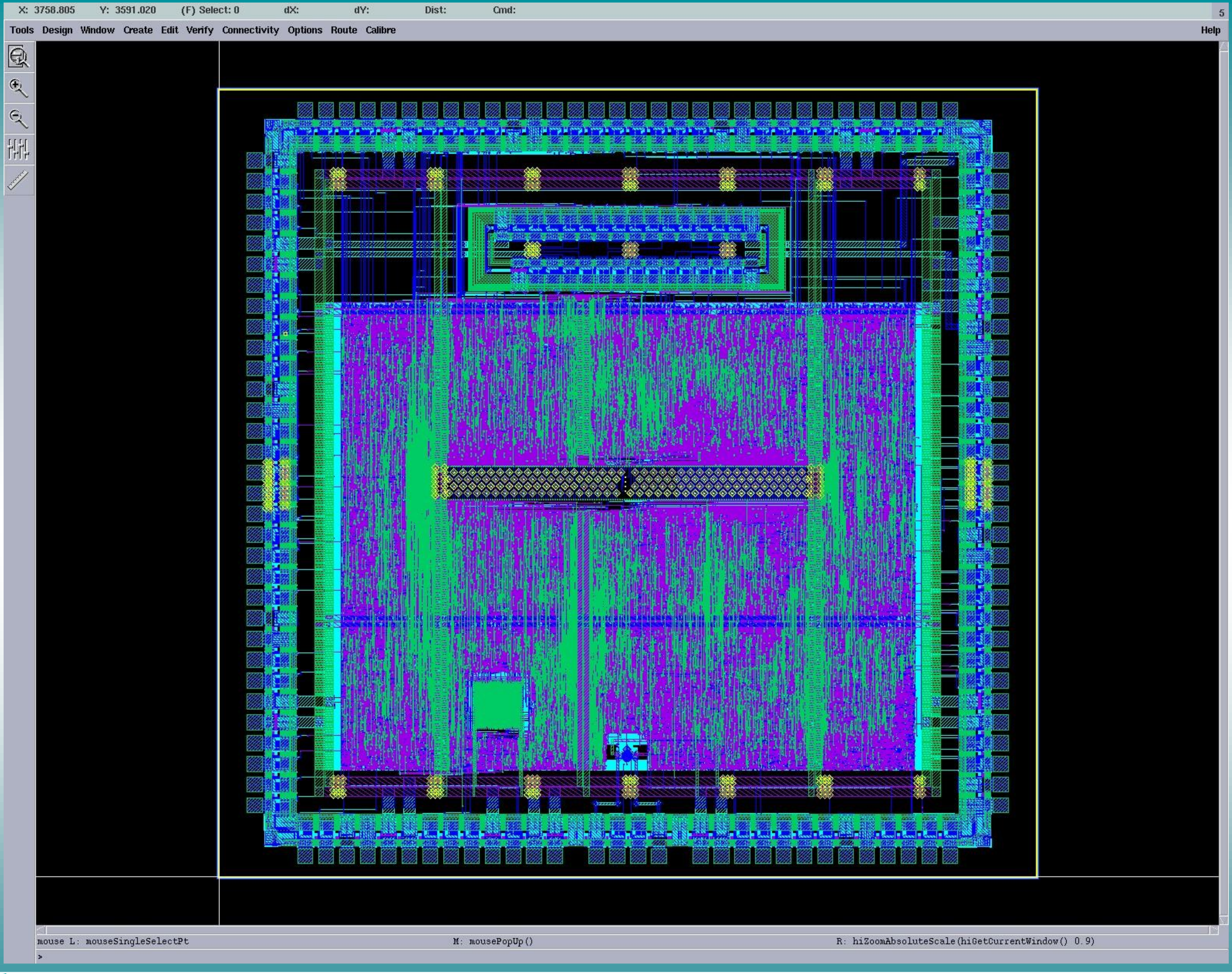
- R8051 CPU
  - 80MHz operation; 140MHz Lab test (VDD High)
  - 220MHz Memory interface
- IEEE 754 Floating point coprocessor
- 32 bit Integer coprocessor
- 2 UARTs, Int. Cont., 3 Timers, …
- Crypto functions
- 128KBytes/layer main memory

- 5X performance
- 1/10th Power

R8051/Memory



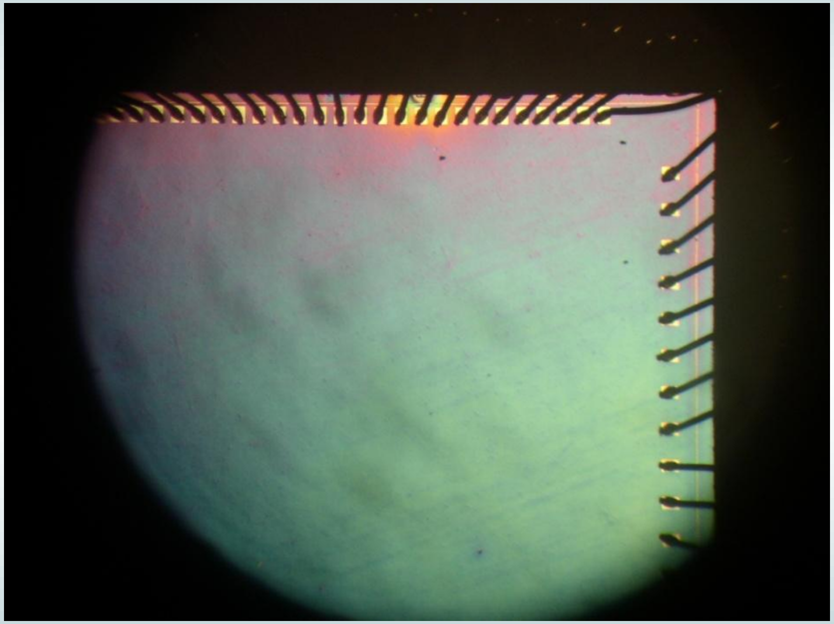


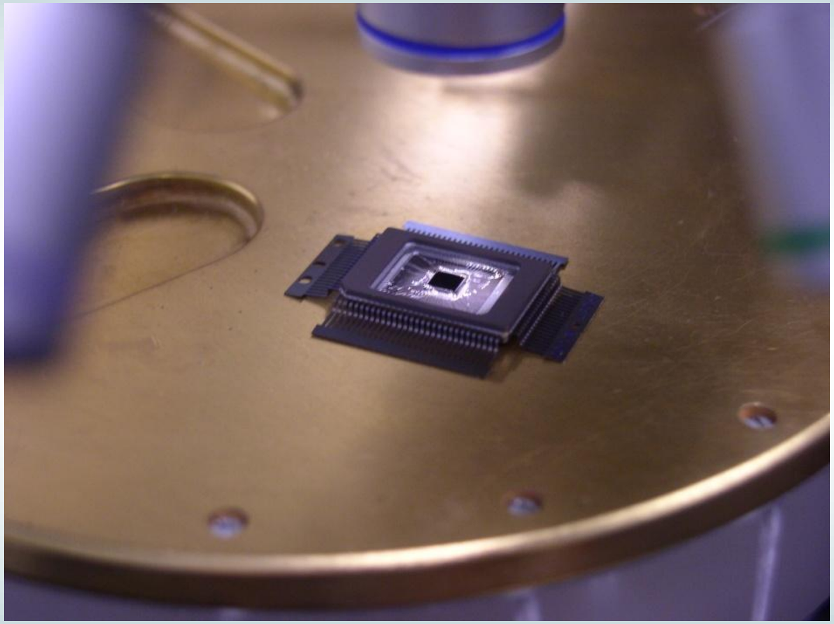

X: 3758.805
Y: 3591.020
(F) Select: 0
dX:
dY:
Dist:
Cmd:
Tools Design Window Create Edit Verify Connectivity Options Route Calibre
Help
mouse L: mouseSingleSelectPt
M: mousePopUp()
R: hiZoomAbsoluteScale(hiGetCurrentWindow() 0.9)

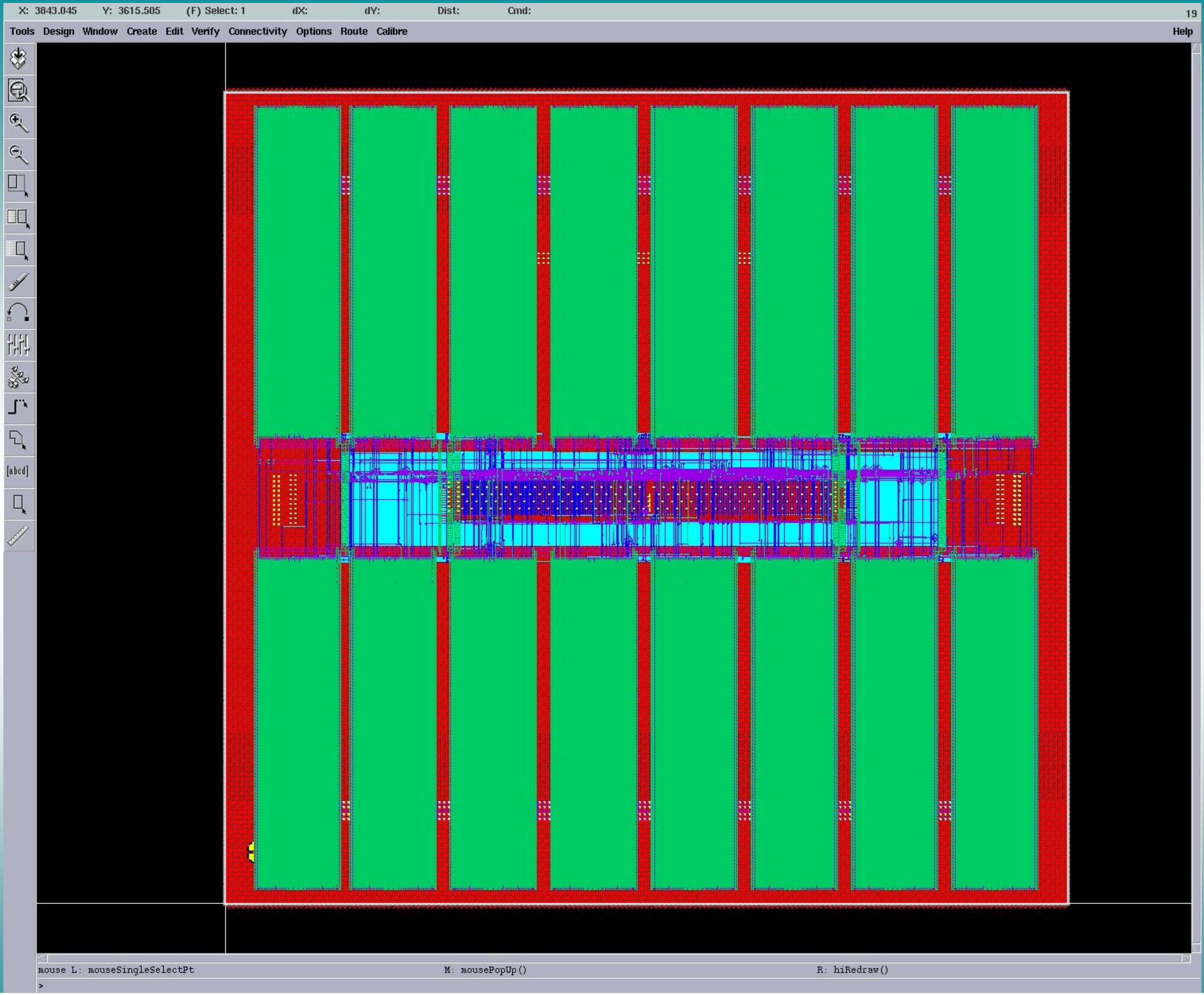
SRAM

- Standard x32 syncburst SRAM
- 128KBytes per memory layer
- Shared layout with R8051 memory
  - Standardized interchangeable 3D blocks

SRAM







X: 3843.045
Y: 3615.505
(F) Select: 1
dX:
dY:
Dist:
Cmd:
19
Tools
Design
Window
Create
Edit
Verify
Connectivity
Options
Route
Calibre
Help
mouse L: mouseSingleSelectPt
M: mousePopUp()
R: hiRedraw()

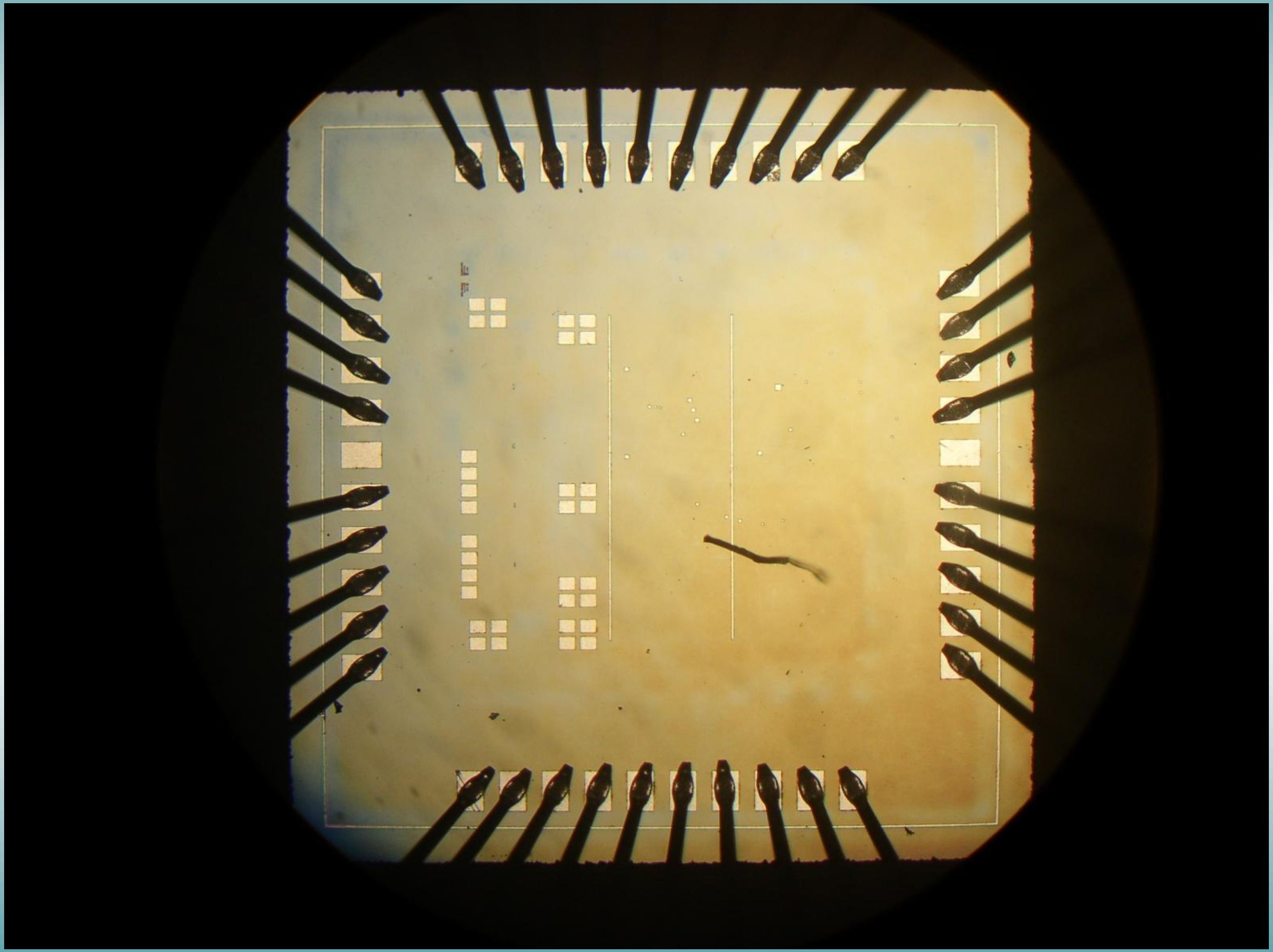
DRAM

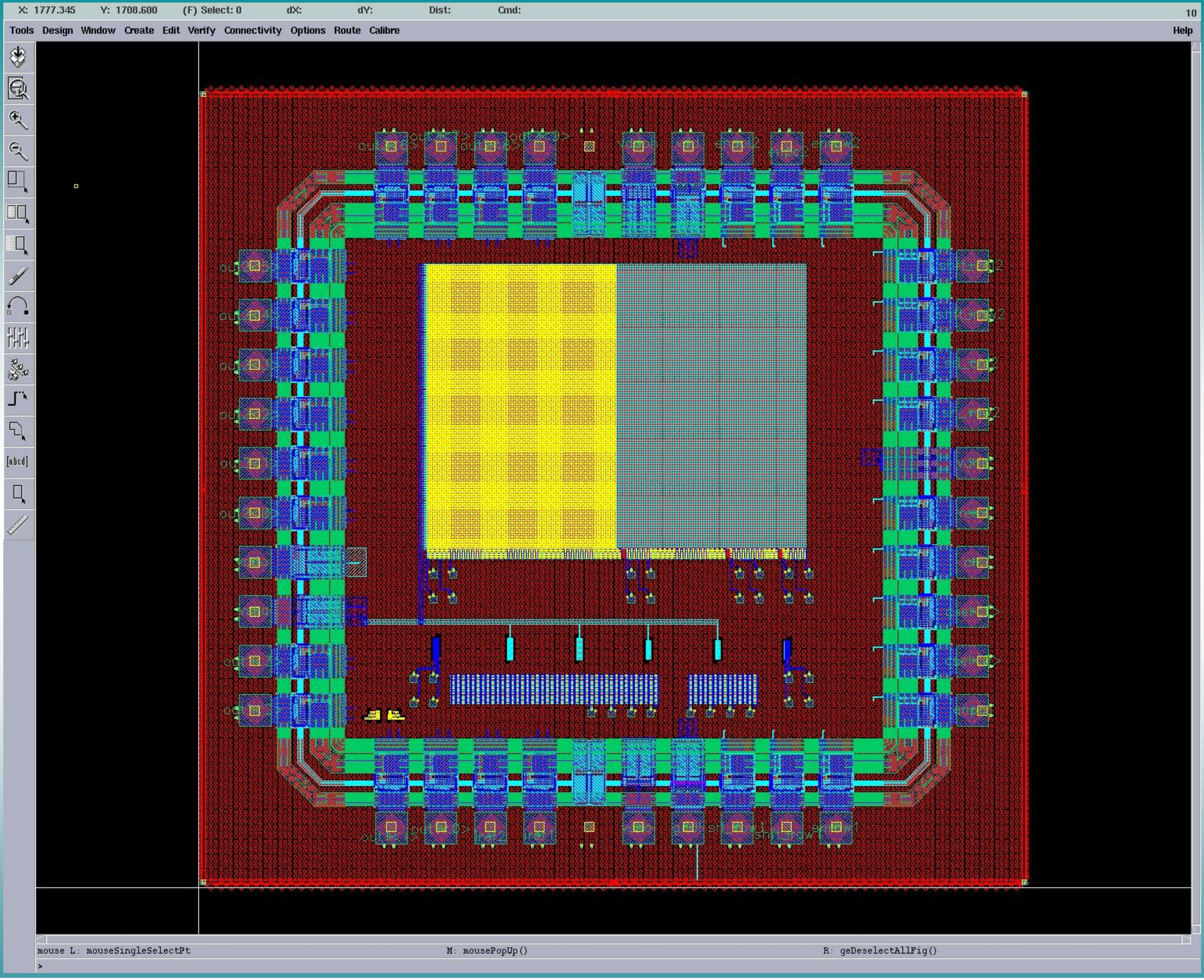
- Split senseamps and drivers from memory cells
- Process separation
- DRAM memory layers, done in <15 Mask layers
- Applicable to flash
- Ideal application, R&R friendly

CMOS Sensor

- Backside illumination
- 100% array efficiency
- 5 pixel fields
  - Various designs
- Main field 160 x 120 pixels; 5x5 um pixel
- Sub-fields 2.4x2.4 and 2.9x2.9 um pixels
- Interconnect @ 2.4um pitch
- Very high sensitivity
- Alignment verification structures

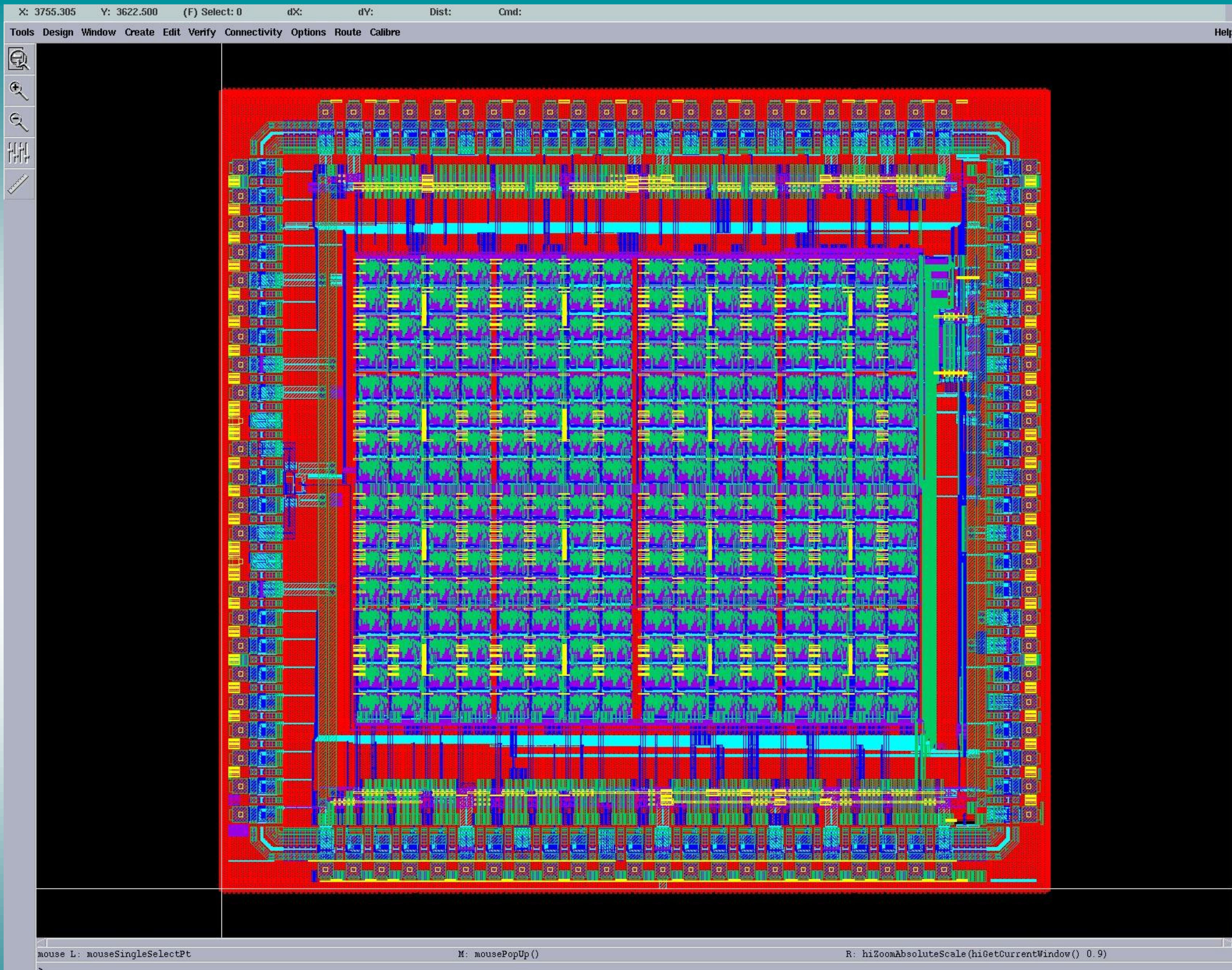
CMOS Sensor




X: 1777.345 Y: 1708.600 (F) Select: 0 dX: dY: Dist: Cmd:
10
Tools Design Window Create Edit Verify Connectivity Options Route Calibre
Help
mouse L: mouseSingleSelectPt
M: mousePopUp()
R: geDeselectAllFig()

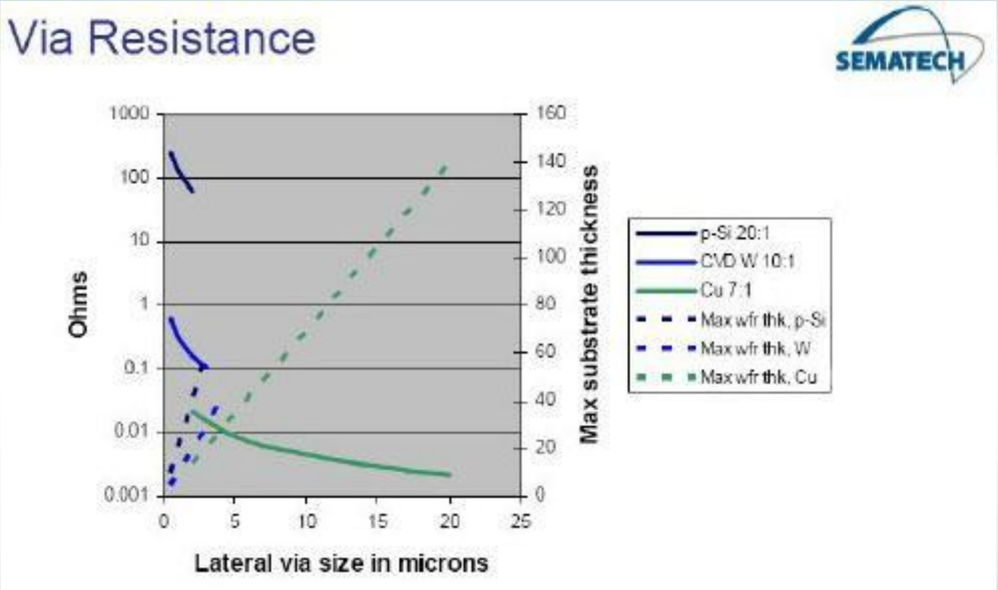
FPGA

- Full 3D interconnect network
- 12 Z axis interconnects per LCA
- 2.5Gb serial link format
- N layer architecture


X: 3755.305 Y: 3622.500 (F) Select: 0 dX: dY: Dist: Cmd:
Tools Design Window Create Edit Verify Connectivity Options Route Calibre Help
mouse L: mouseSingleSelectPt
M: mousePopUp()
R: hiZoomAbsoluteScale(hiGetCurrentWindow() 0.9)

Boundaries

- Z dimension increments
  - 5-15um thickness for wafer and chip on wafer
  - 50um for chip on chip
- Low R
- Very Low to Moderate C
  - 3ff for wafer to wafer
  - 25ff for chip to wafer
  - 1-5pf for chip to chip
- Repair & Redundancy
  - It's still per sqmm!
- Pitch
  - 0.5um limit for wafer level
  - 10um for chip on chip
- How many layers?
  - 2 to 5, current horizon for wafer level
  - 3+ chip on wafer
  - 8+ chip stacking



HEAT!!!

- Modeling
  - Lots of modeling, need more real data, more testing required…
    - Seems the limit is governed by TIM, ~1W/sqmm
- What we know….
  - 32W/sqmm, Structurally sound
- <5W easy rules
  - ~15W/100sqmm cliff
  - >150W possible
  - >500W liquid cooling

3D Problems

- Stacking reduces yields:
  - –NO!
  - –3D interconnect failure <0.1ppm
  - –Yield primarily is a function of cumulative die size
- Memory, die on wafer has KGD issues;
- Was the die good?
  - –At speed probe
  - –1000's of I/O
  - –Probe damage
- Is it good after processing?
  - –Tezzaron memory is post attachment repairable
  - –Self test drastically reduces probe
  - –Failure isolation for FA
- Supply Chain
  - –Foundries

Costs

- TSV overhead
    - 0 to <10%
- Processing
    - – TSV cost
    - – Bonding
    - – Thinning
        - High volume <$150 per bond… can see <$25 future
- Yield
    - Still primarily by sqmm, but could be better or worse
- Prototyping
    - – +2-3 masks (large geometry)
    - – Complete maskset per layer (could do up to 4 on 1 for MPW)
    - – ~$25K Lot tooling charge, $2,500 per wafer

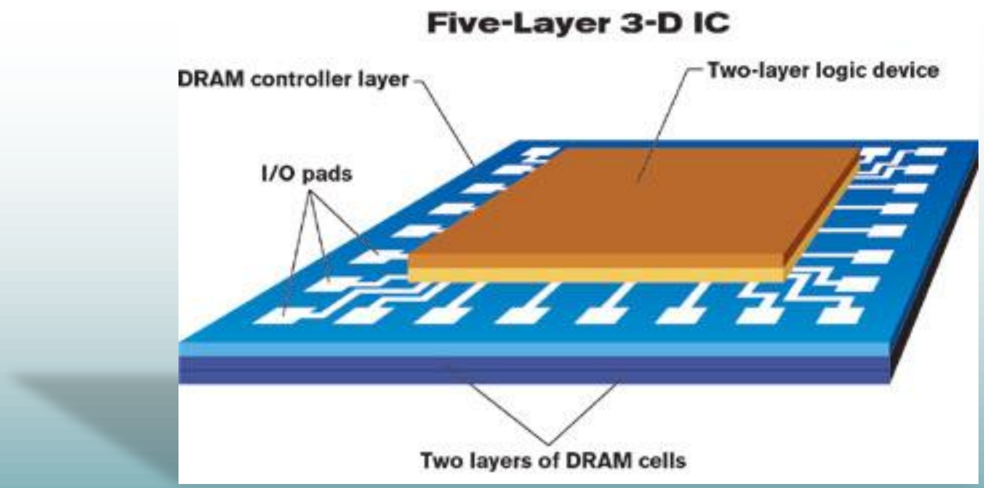
Qual data

- 100,000 device temperature cycles -65/+150
  - No failures
  - Two build lots
- 168 hour high temp
  - No Failures
  - Extended to 336 and then 504 with no failures
- Hot spot delamination testing
  - >10watts/sqmm, no failures
- Life test under bias
  - >10,000 hours, no failure
- 3D interconnect failure ~100ppb

Current Developments

- MPW Runs
- Memory to logic
- 8 layers
- Rad hard
- Expect 30-50 different completed devices in 09.

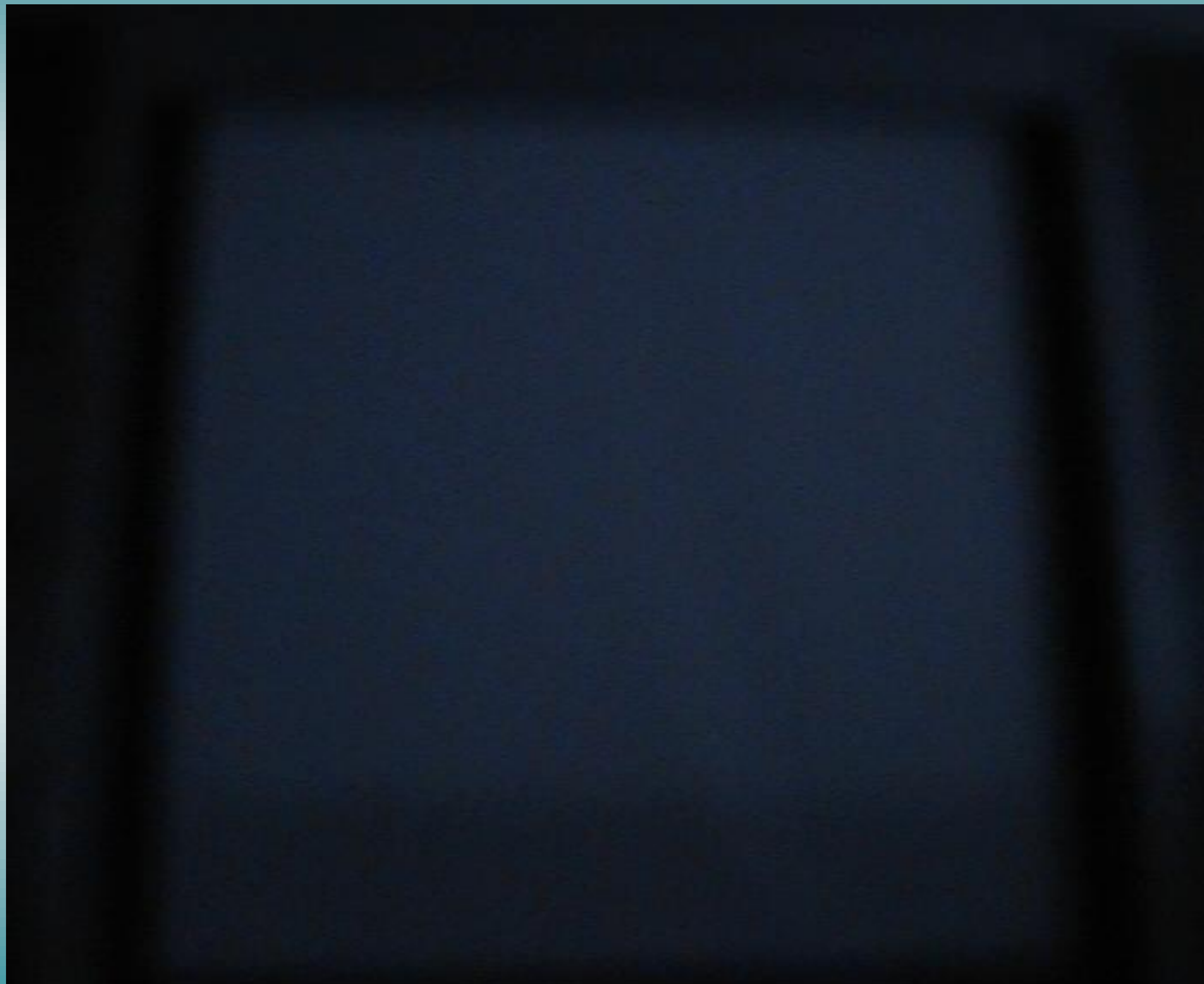
Proposed 5 layer stacks



2009 MPW Schedule

- 130/110nm Memory centric 2Q09
  - Wafer-wafer and chip-wafer
  - 6 + Tezzaron
- 130nm Mixed signal 2Q09
  - Wafer-wafer
  - ~12 participants
- 130/110nm Memory centric 4Q09
  - Wafer-wafer
- 130nm with memory, 2 to 5 total layers 4Q09
  - Wafer-wafer and chip to wafer (Stack of stacks)
  - R3Logic/NC State
  - 18 participants
- 90nm Mixed signal, HBD, HPC 4Q09-1Q10
  - Wafer-wafer and chip to wafer

2D/3D Processor Race



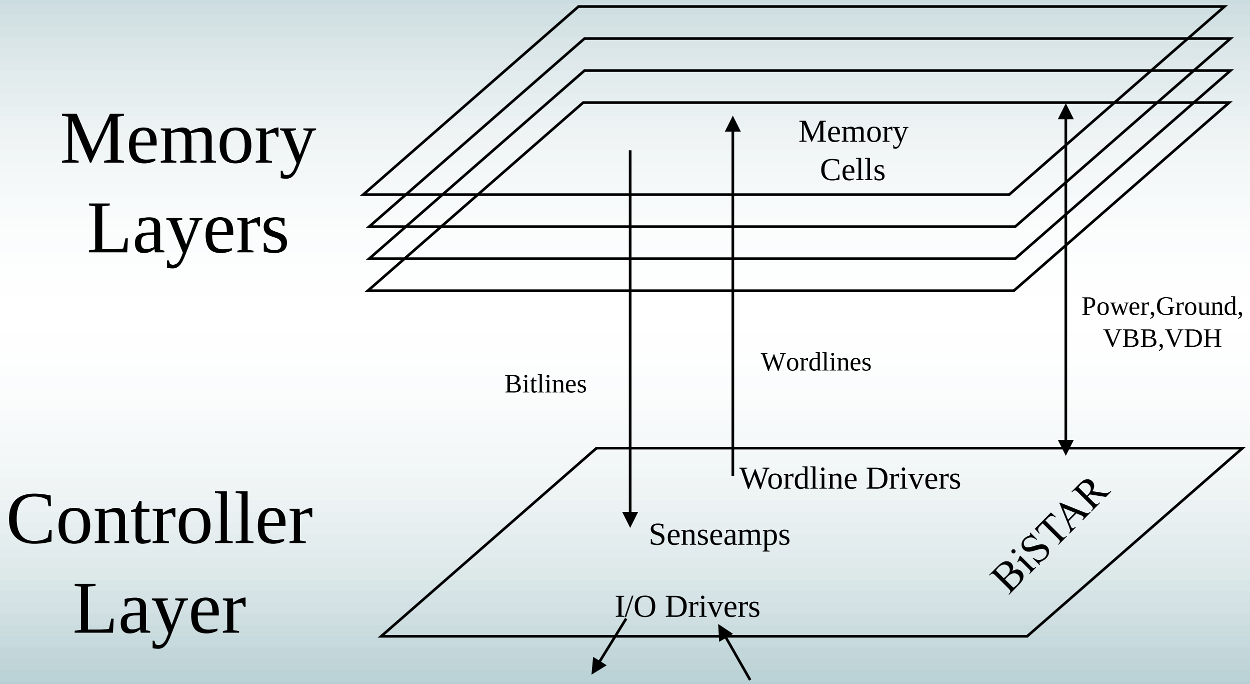
What can 3D DRAM achieve?

- Faster Access Time
- Lower Power
- Denser
- Reliable
- Compatible
- Lower Costs

DRAM wants 2 different processes!

Bit cells	Low leakage -slow refresh -low power -low GIDL	High Vt Devices Vneg Well Thick Oxide
Sense Amps Word line drivers Device I/O	High speed -better sensitivity -better bandwidth -lower voltage	Low Vt Devices Copper interconnect Thin Oxides

“Dis-Integrated” 3D Memory



Memory Layer

- The memory layer contains 128Mb tiles which can be stacked vertically and/or used in any 2D configuration.
  - So one project might be using two memory layers, each being a 2 x 2 set of 128Mb tiles; Another project could be using four layers of memory, each layer containing a 2 x 4 set of 128Mb tiles.
    - These use the same memory layer! The customization is only in the Controller layer. This greatly simplifies production and mitigates risks.

Controller Layer

- The Controller layer can be thought of as giving the memory its personality.
  - Various pre-designed pieces, such as sense-amp blocks, word line drivers, BIST components are pulled together to enable basic memory function.
  - Additional custom functions can be added, as well as accommodating the required unique I/O parametrics and functionality.

❖ Due to the intimate attachment and the logic process used for this layer, I/O not only can be very wide, but also run faster than 4GT/s. Also no address or data coding, such as that used in GDDR memories is required.

Increasing Die Overhead

Array Utilization

DDRI 70%

DDRII 47%

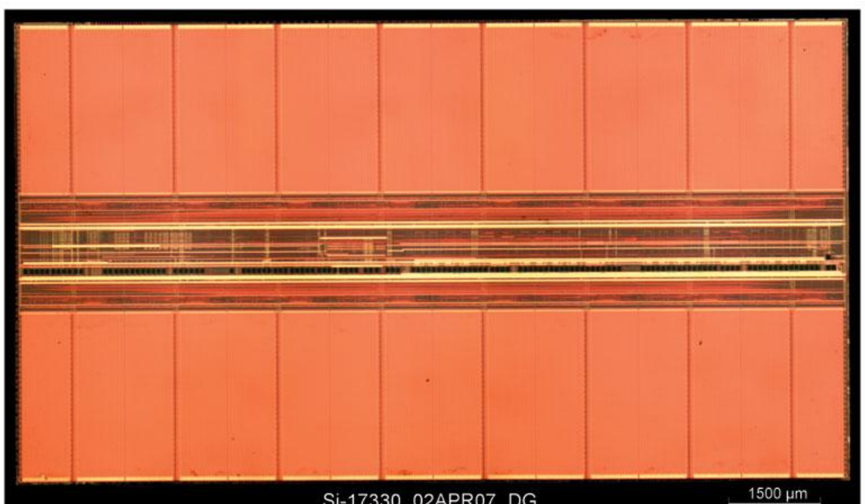
DDRIII 38%

DDRIV <30%?

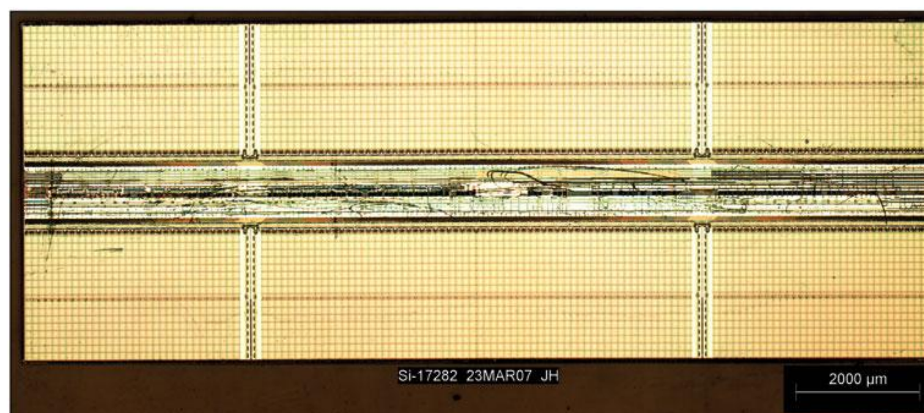
Chip size overhead of DDR3 relative to DDR2

	90 nm		80 nm	
Device density	DDR2	DDR3	DDR2	DDR3
Chip size	1	1.22	1	1.23
Gross dice per wafer	1	0.81	1	0.82

Source: Semiconductor Insights

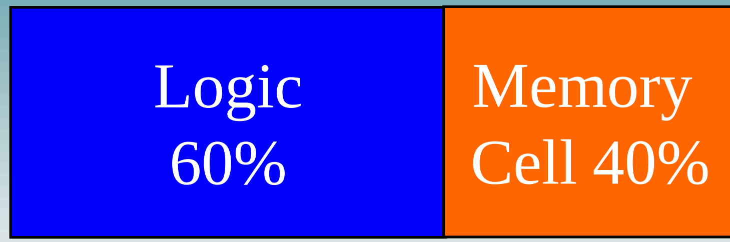


Die photo of Micron's 1-Gbit DDR3, which has a gross-dice estimate of 600 per 300-mm wafer in 78-nm technology.

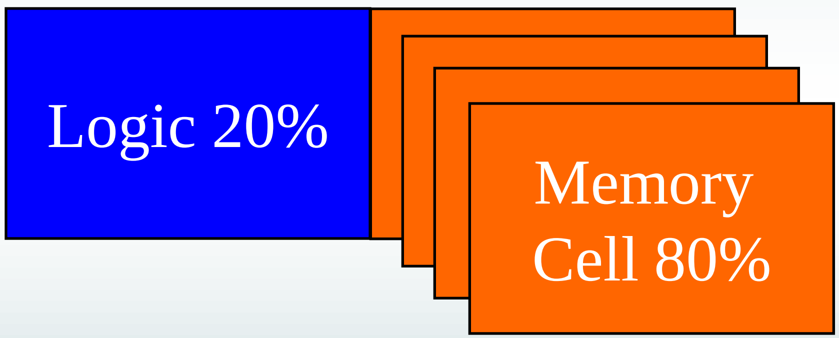


Die photo of Qimonda's 512-Mbit DDR3, which, at 14.6 x 5.4 mm, is more of a rectangle than the squarish DDR2.

Standard DRAM Utilization



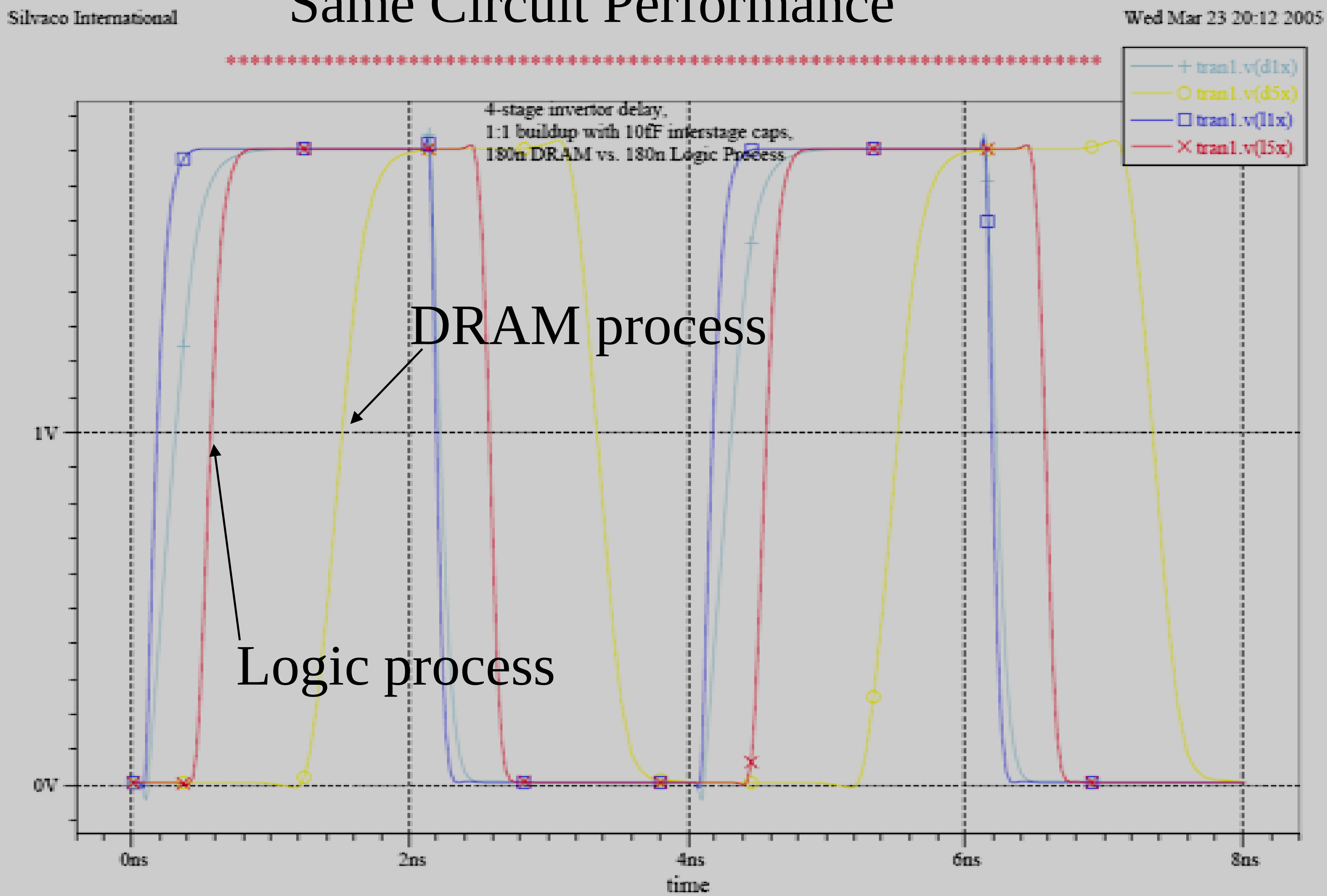
66% Savings in logic per memory cell

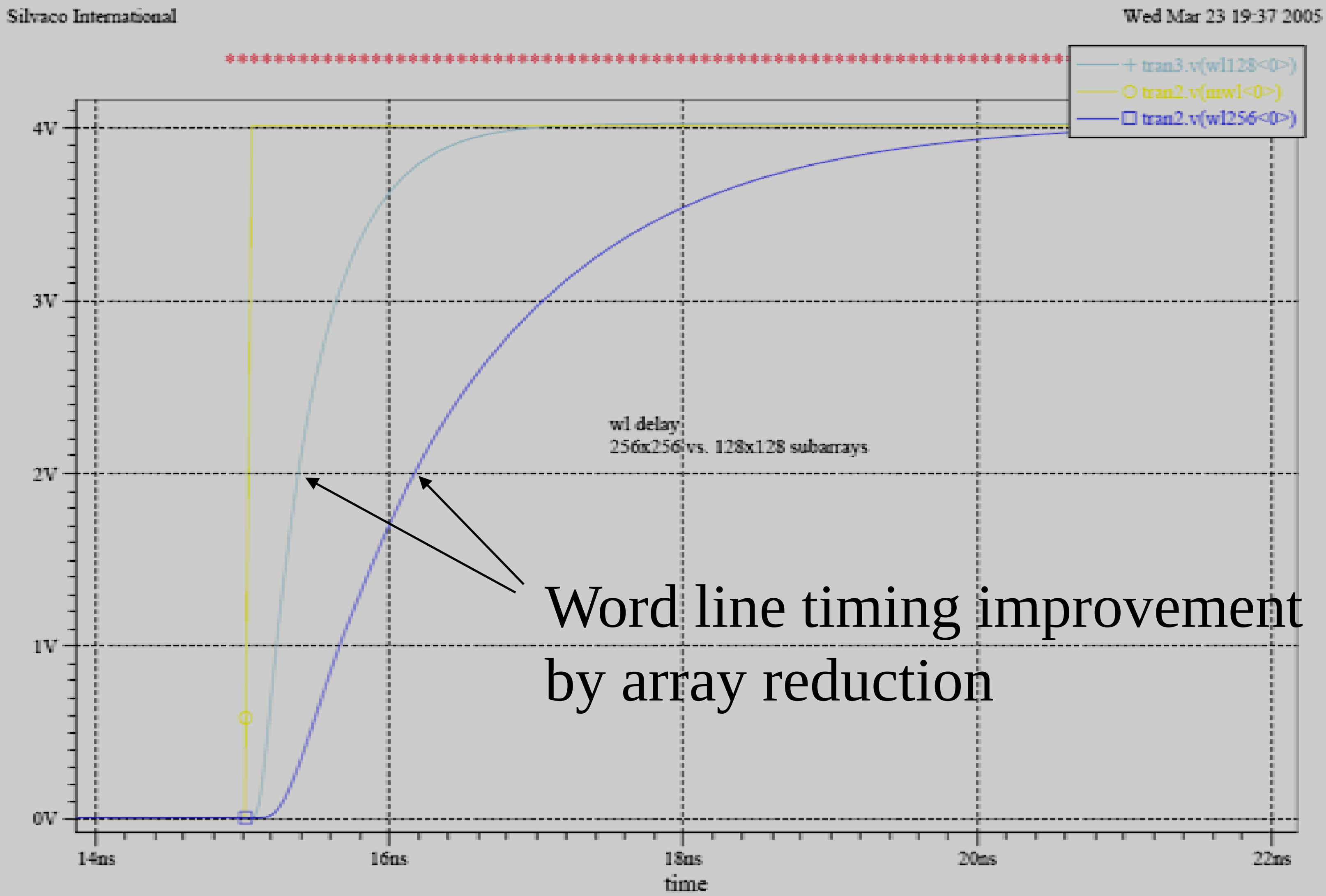


But…..this requires,

Millions of *vertical* interconnect!

Same Circuit Performance

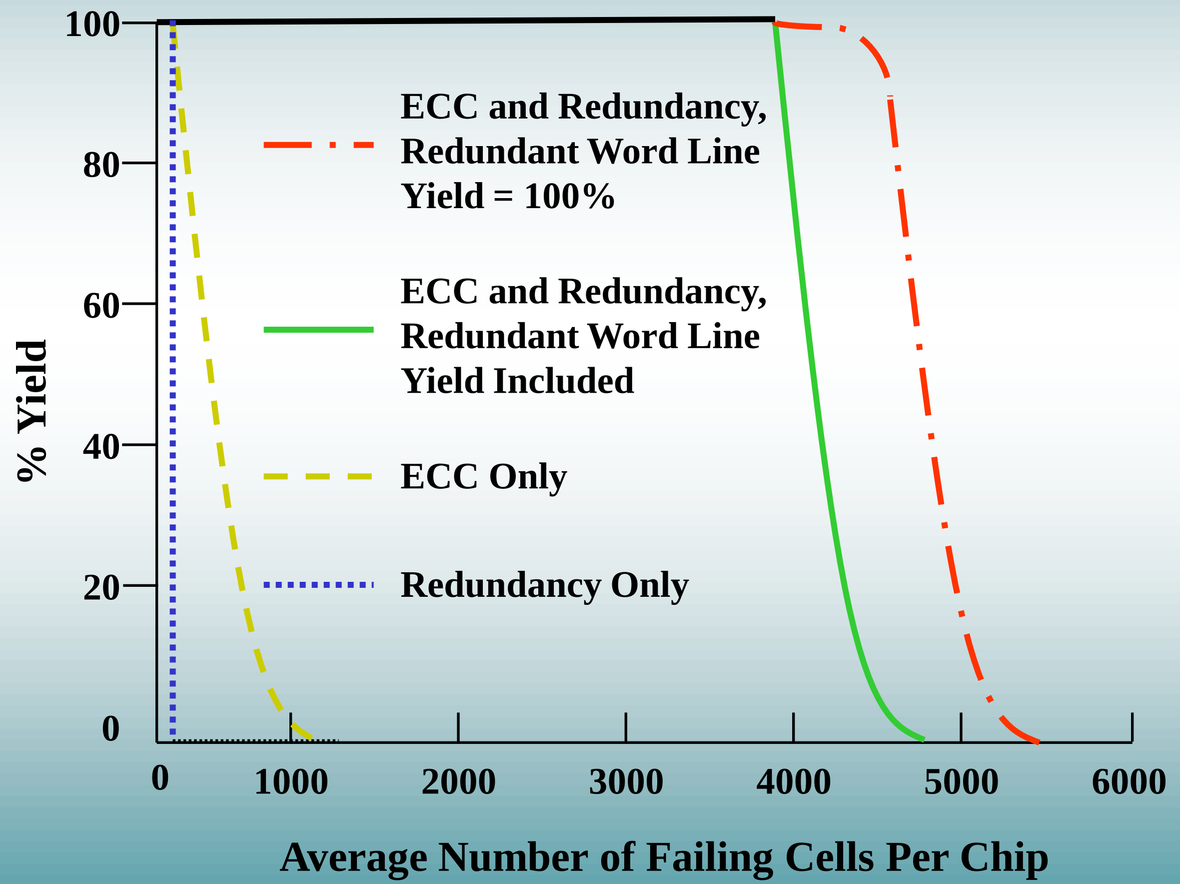



Silvaco International
Wed Mar 23 19:37 2005
tran3.v(wl128<0>)
tran2.v(mwl<0>)
tran2.v(wl256<0>)
4V
3V
2V
1V
0V
14ns
16ns
18ns
20ns
22ns
time
wl delay
256x256 vs. 128x128 subarrays
Word line timing improvement
by array reduction

Tezzaron's Solution to Yield

Tezzaron's IP Bi-STAR™ delivers 97%+ yields = Near doubling of fab capacity

Tezzaron's wafers can tolerate higher numbers of failed cells because Tezzaron's IP enables fine grain cell-by-cell re-mapping, rather than whole row or column trimming, to make repairs.



Bi-STAR™ Technology

- Innovative way to improve the yield of highly parallel structures such as memory
- Basis: integration of intelligent self test, self repair
- Performs greater level of testing than normally available during normal chip or wafer level testing
- Bi-STAR™ tests and compares >300,000 nodes or bits/clock cycle; more than 1,000 times faster than can be achieved by any external memory tester

What Can Bi-STAR™ Test & Repair?

- Bad memory cells
- Bad line drivers
- Bad sense amps
- Shorted word lines
- Shorted bitlines
- Leaky bits
- Bad secondary bus drivers
- Bad CAMS

Superior SEU Tolerance

- Short bit lines
- Thin substrates
- High bit to bitline capacitance ratio
- Onboard ECC
- Full memory scrub every 2 minutes
- On the fly redundancy
- Soft error FIT rate improvement by >10,000x

Lower Costs!

- Less processing per wafer
  - >35% lower processing costs
- Higher array utilization
  - +50%
- Lower test cost (using Bi-STAR™)
  - 95% reduction in test cost
- Higher yield (using Bi-STAR™)
  - +5 to +75%

Octopus Cache DRAM

- 128Mb per layer, per 16sqmm
  - 128Mb to >2GB
- Down to 5ns, latency
- 2GHz Max clock rate
- Minimum Timing - tRCD=1, tCYC=4, tPRE=0, tCL=2
- Programmable burst length 4 to 256
- Programmable port width 32 to 256 bits
- Exposed or hidden refresh options
- DDR 4000MT Max
- Internally ECC protected, Dynamic self-repair, Post attach repair
- 115C die full function operating temperature
- For 8 port 1-4Gb core
  - >200GB/s *sustained,* closed page mode, BL=4, bandwidth
  - 1TB/s peak bandwidth
  - >25TB/s peak on-board transfer rate